Chiplet-based electronics: evolution or revolution?

Alberto Sangiovanni-Vincentelli

The Edgar L. and Harold H. Buttner Chair of EECS,
University of California at Berkeley

Co-founder and Member of the Board,
Cadence

© 2024 Alberto Sangiovanni Vincentelli

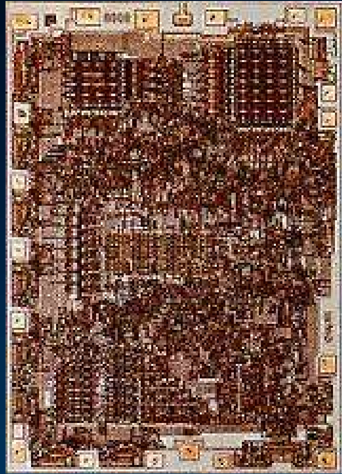
Berkeley UNIVERSITY OF CALIFORNIA cadence

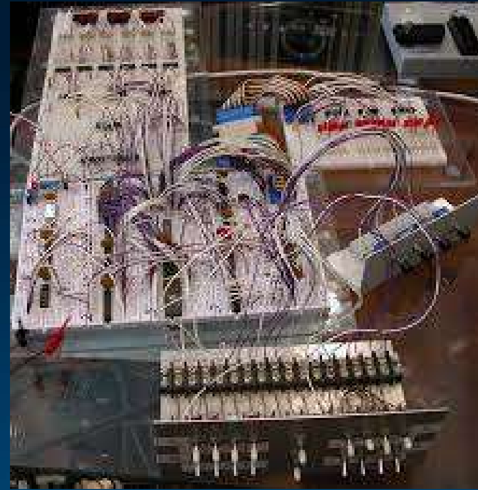
Outline

- Big chips or chiplet-based implementation?
- Early Chiplet Examples
- The Design Problem
- Is AI a Panacea?
- Design Flows for Chiplet-based design
- A Bit of Research

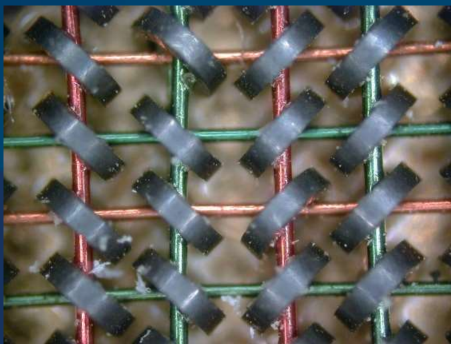
© 2024 Alberto Sangiovanni Vincentelli

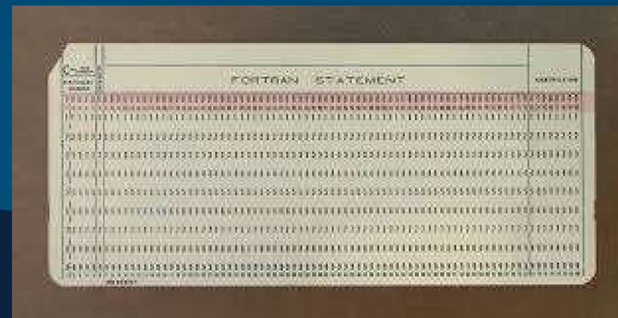
Once upon a time….


3,500 Transistors, 92K IPS


Verifying the chip

Preparing the masks

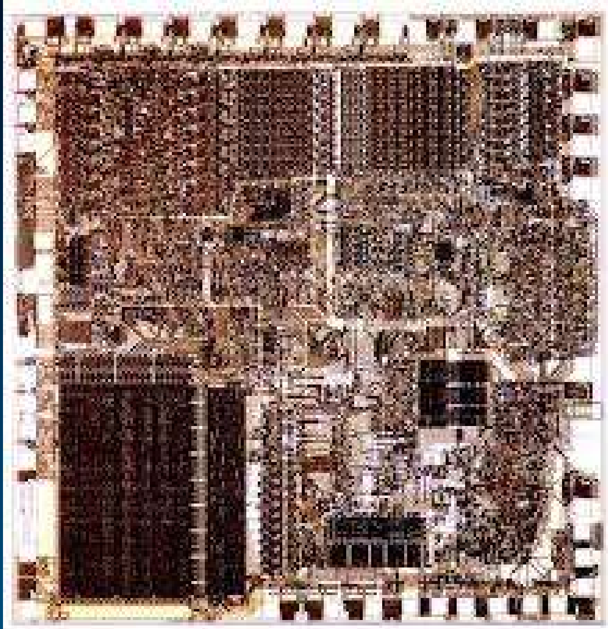

Magnetic core memories



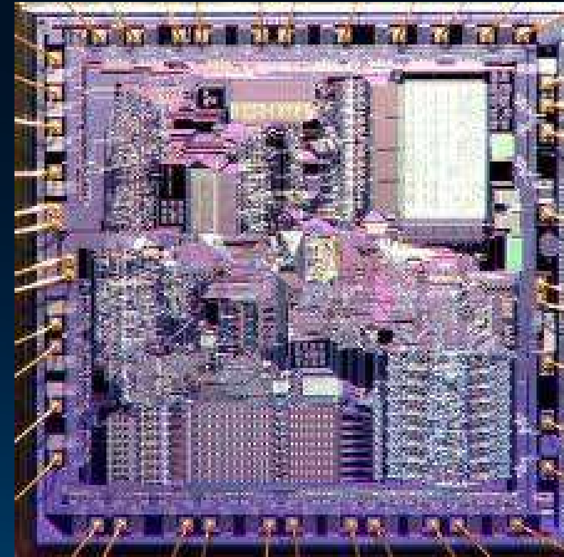
Punched Card

© 2024 Alberto Sangiovanni Vincentelli

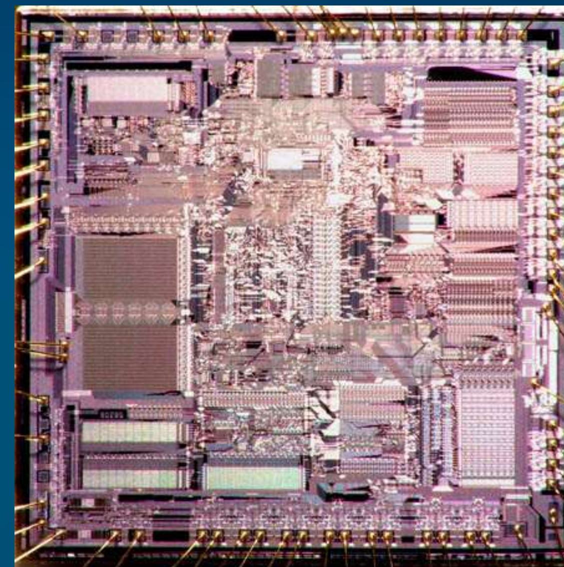
Evolution: From Handcraft to...



Intel 4004



Intel 8086

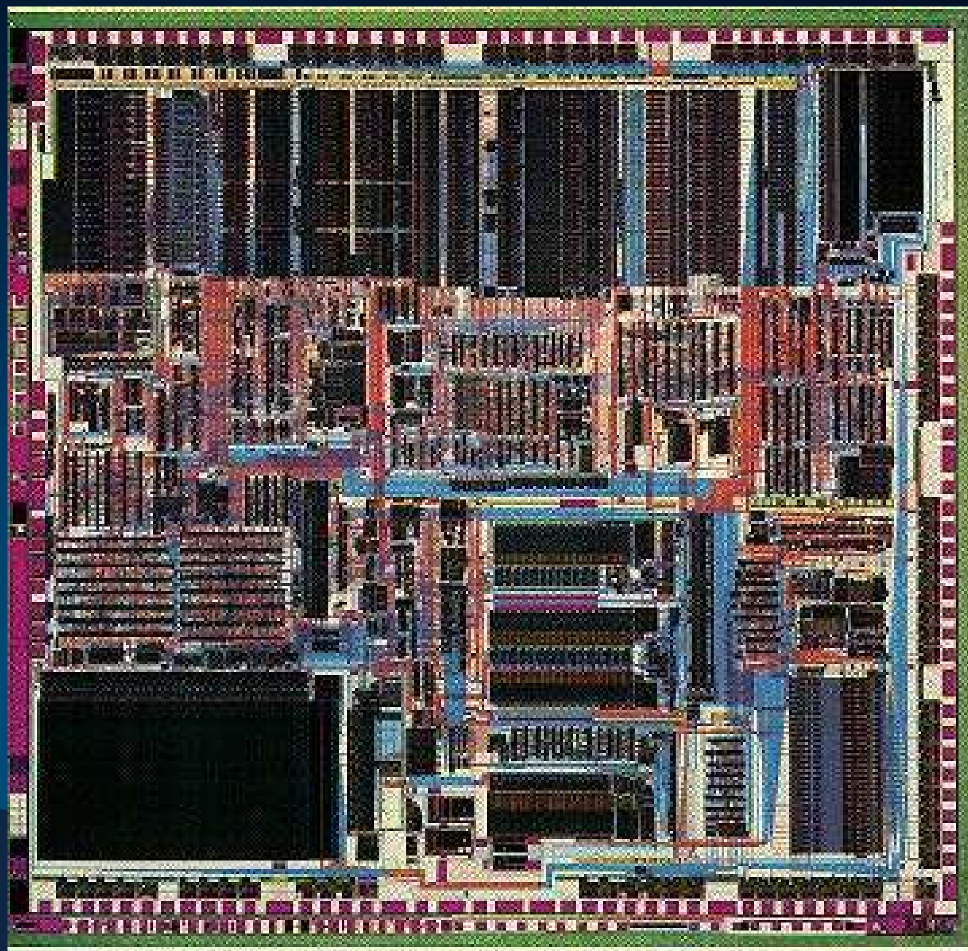


Intel 80286

© 2023 Alberto Sangiovanni Vincentelli

Regularity, Methodology, Re-usable Parts and Tools!

https://vcresearch.berkeley.edu/news/history-innovation-berkeley-entrepreneurs-companies-changed-way-we-live



© 2024 Alberto Sangiovanni Vincentelli

Berkeley
UNIVERSITY OF CALIFORNIA

1983: Intel 386

© 2023 Alberto Sangiovanni Vincentelli

Engineering SCIENCE

Segesta (Σέγεστα) Temple, Sicily, 420 BC (Picture Taken 2020)

© 2024 Alberto Sangiovanni Vincentelli
Berkeley
UNIVERSITY OF CALIFORNIA

How did we cope with complexity?

(ASV, Corsi e Ricorsi: The EDA Story, IEEE Solid State Circuits Magazine, 2010)

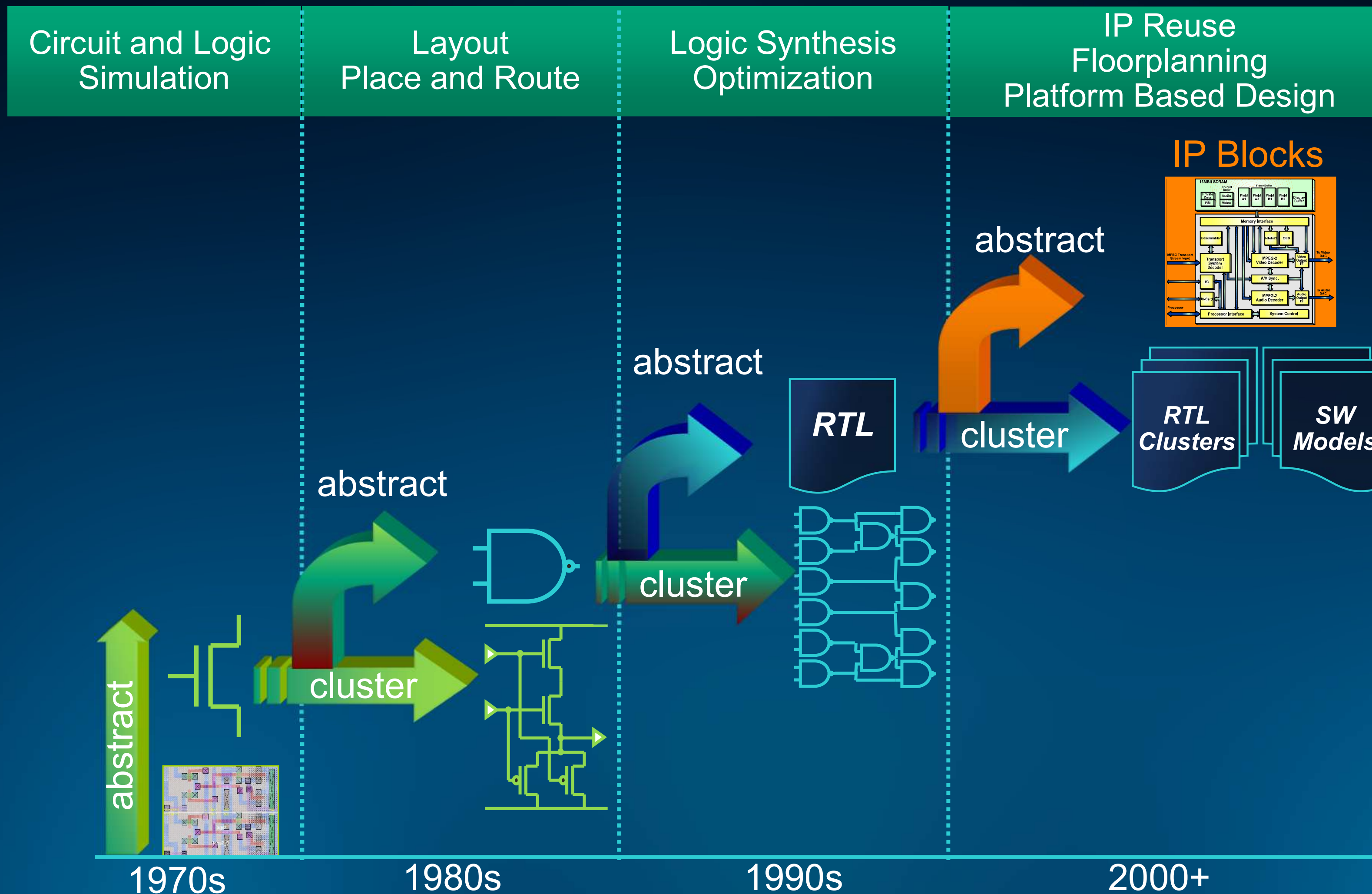
Abstractions

Methodologies
(Freedom from choice)

Tools

© 2024 Alberto Sangiovanni Vincentelli

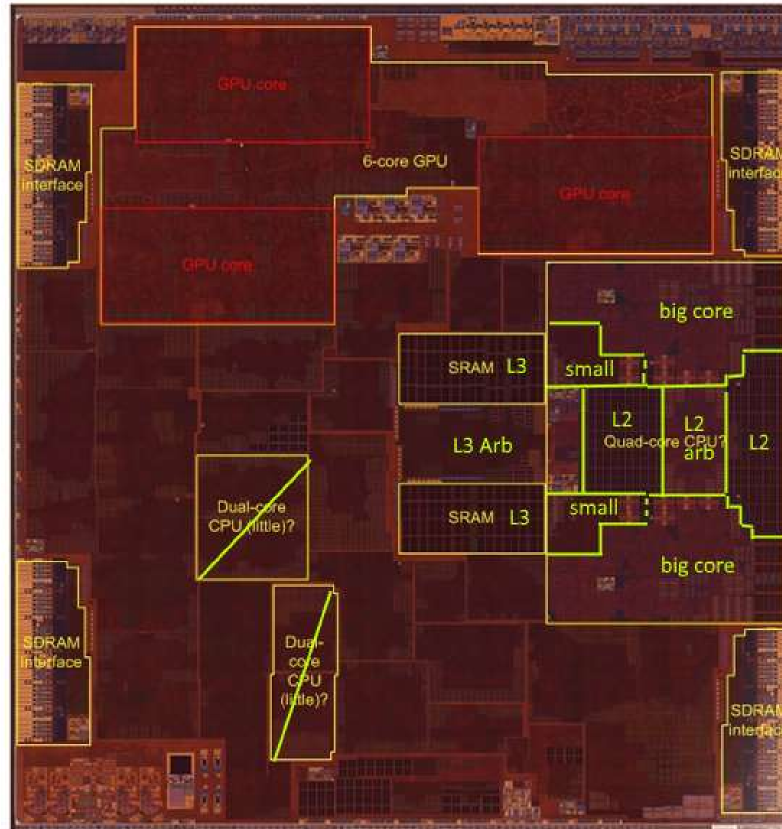
EDA and Design Methodology Evolution: What is Next?



© 2023 Alberto Sangiovanni Vincentelli
Berkeley UNIVERSITY OF CALIFORNIA cadence

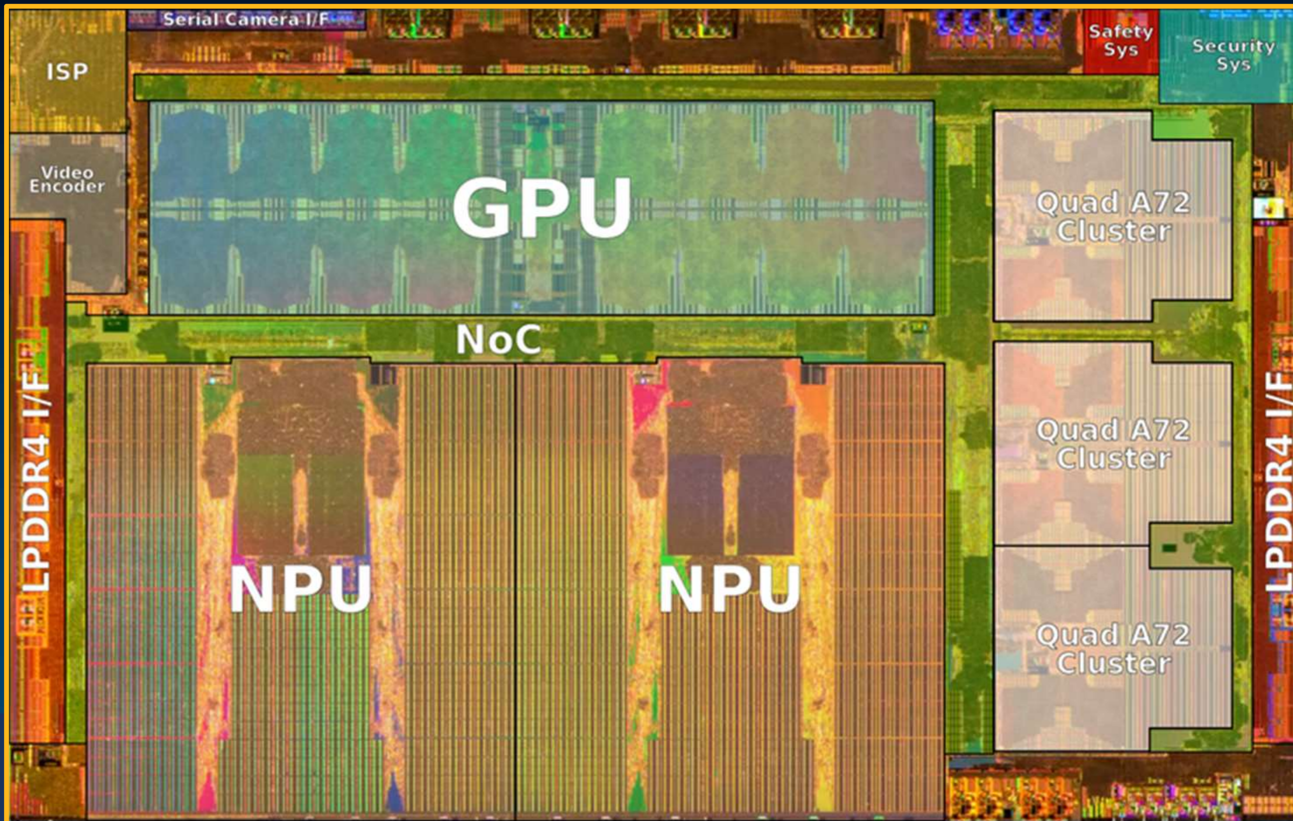
Apple Monster Chip

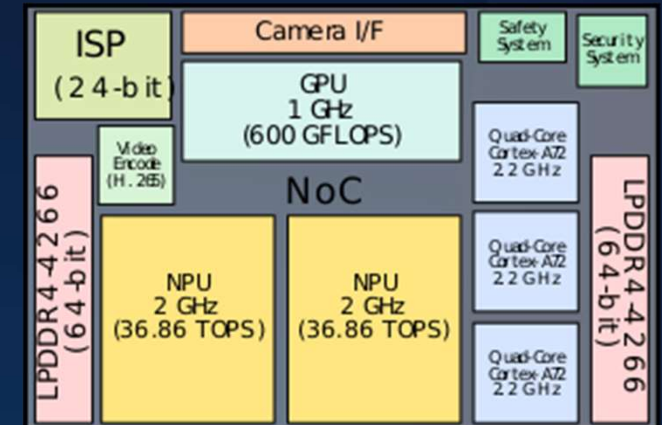
Apple A11 Bionic
4.3Billion transistors
34 GOPS
87.66 mm2



© 2024 Alberto Sangiovanni Vincentelli
Berkeley
UNIVERSITY OF CALIFORNIA

TESLA FSD SoC (36.8 teraOps, 260mm^2 , 6Billion Tr)



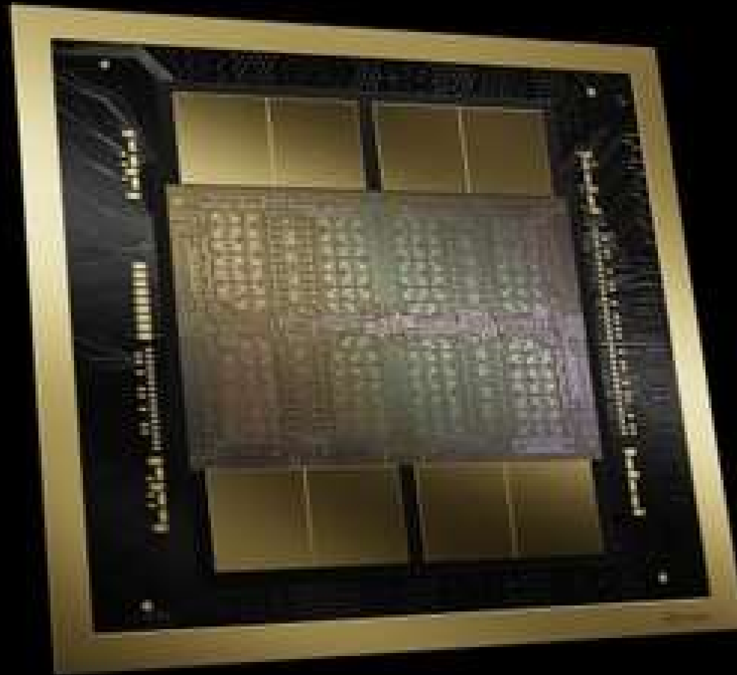


- NoC – Network on chip
- ISP – Image Signal processing
- Safety Sys – Lock step for ISO26262
- Security – only TESLA certified software

Chip focused on Automotive L5 use case for Deep learning

© 2024 Alberto Sangiovanni Vincentelli

Nvidia Blackwell SoC 208 billion transistors



© 2023 Alberto Sangiovanni Vincentelli

© 2024 Alberto Sangiovanni Vincentelli

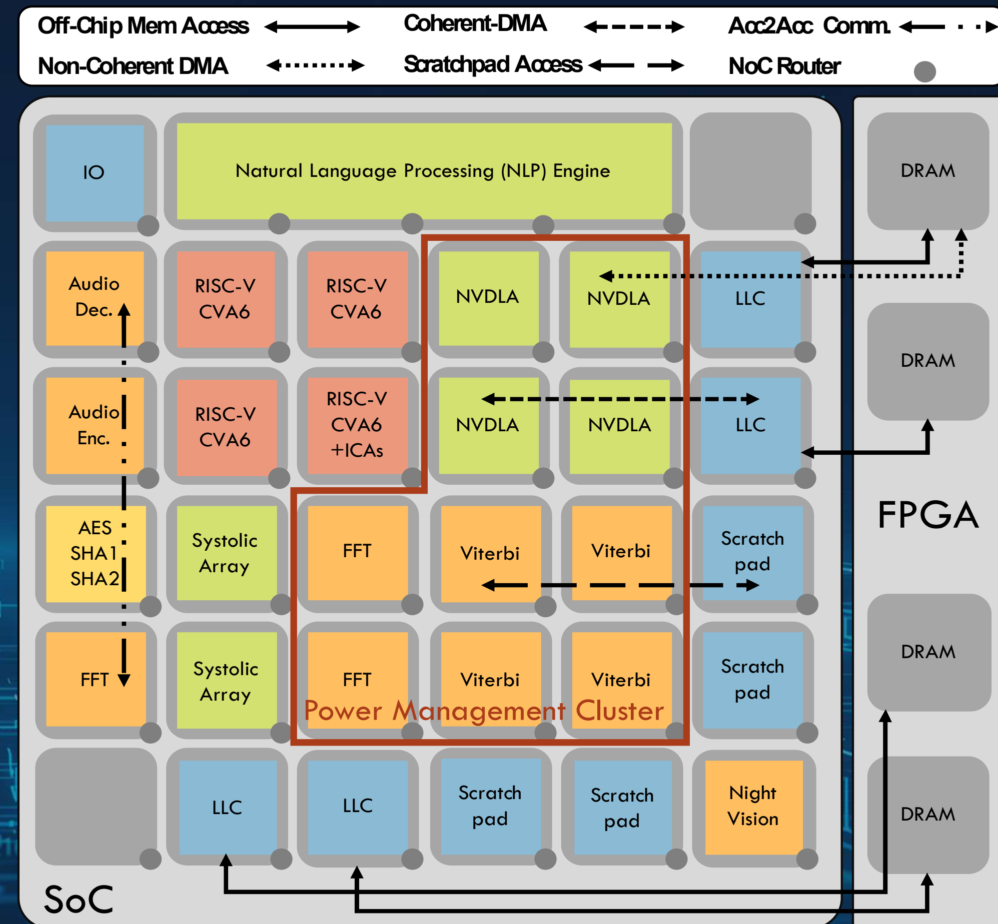
The Concept of SoC Platform

- **Innovation in SoC architectures and their design methodologies needed to promote design reuse and collaboration**
  - Architectures and methodologies must be developed together
- ***Platform = architecture + methodology***
  - An SoC architecture enables design reuse when it simplifies the integration of many components that are independently developed
  - An SoC methodology enables design collaboration when it allows designers to choose the preferred specification languages and design flows for the various components
- **An effective combination of architecture and methodology is a platform that maximizes the potential of open-source hardware**
  - by scaling-up the number of components that can be integrated in an SoC and by enhancing the productivity of the designers who develop and use them

© 2024 Alberto Sangiovanni Vincentelli

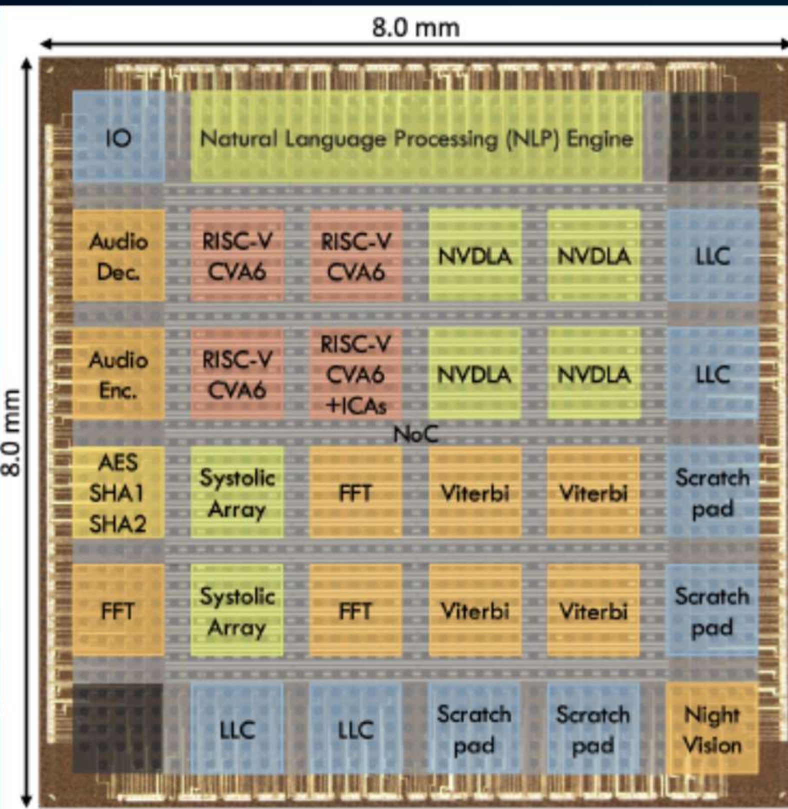
The EPOCHS-1 SoC: Chip Highlights

- 64 mm^2 SoC designed in 12 nm FinFET
- 35 clock domains; 23 power domains
- 8.4 MB on-chip SRAM memory
- Tile-based SoC architecture
- 34 tiles connected by a 6-plane 2-D mesh NoC
- The 74 Tbps NoC provides flexible orchestration of data
- 4 RISC-V processor tiles booting Linux SMP
- 23 accelerators of 14 different types
- 10 accelerators compose a cluster demonstrating a novel distributed hardware power management scheme
- Designed by a small team of PhD students, postdocs, and industry researchers in 3 months with ESP, the open-source platform for agile SoC design



© 2024 Alberto Sangiovanni Vincentelli
Berkeley UNIVERSITY OF CALIFORNIA
cadence

ESP is Silicon Proven: The EPOCHS-1 SOC



Technology	12nm FinFET
Area	$64mm^2$
#IOs	340
Power Domains	23
Clock Domains	35
Power	83mW – 4.33W
Total SRAM	8.4MB
Max Frequency Range	680MHz – 1.6GHz
Example Application Domain	Collaborative Autonomous Vehicles

14.5 A 12nm Linux-SMP-Capable RISC-V SoC with 14 Accelerator Types, Distributed Hardware Power Management and Flexible NoC-Based Data Orchestration

Maico Cassel dos Santos*[1], Tianyu Jia*[2], Joseph Zuckerman*[1], Martin Cochet*[3], Davide Giri[1], Erik Jens Loscalzo[1], Karthik Swaminathan[3], Thierry Tambe[2], Jeff Jun Zhang[2], Alper Buyuktosunoglu[3], Kuan-Lin Chiu[1], Giuseppe Di Guglielmo[1], Paolo Mantovani[1], Luca Piccolboni[1], Gabriele Tombesi[1], David Trilla[3], John-David Wellman[3], En-Yu Yang[2], Aporva Amarnath[3], Ying Jing[4], Bakshree Mishra[4], Joshua Park[2], Vignesh Suresh[4], Sarita Adve[4], Pradip Bose[3], David Brooks[2], Luca P. Carloni[1], Kenneth L. Shepard[1], Gu-Yeon Wei[2]

[1]Columbia University, New York, NY; [2]Harvard University, Cambridge, MA
[3]IBM Research, Yorktown Heights, NY; [4]University of Illinois, Urbana, IL
*Equally Credited Authors

16

ISSCC 2024 / SESSION 14 / DIGITAL TECHNIQUES FOR SYSTEM ADAPTATION, POWER MANAGEMENT AND CLOCKING / 14.5

© 2023 Alberto Sangiovanni Vincentelli
Berkeley UNIVERSITY OF CALIFORNIA cādence

Luca Benini Group (Bologna+ETH):

Multiple architectures with heterogeneous accelerators

© 2024 Alberto Sangiovanni Vincentelli

Start Small: Open Platform for Autonomous Nano-Drones

Advanced autonomous drone

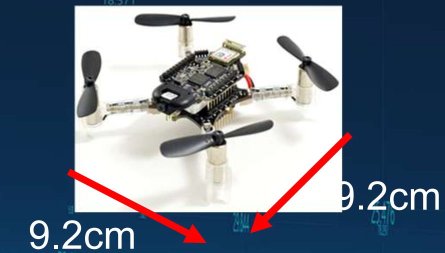
[1] A. Bachrach, “Skydio autonomy engine: Enabling the next generation of autonomous flight,” IEEE Hot Chips 33 Symposium (HCS), 2021

https://www.skydio.com/skydio-2-plus

- 3D Mapping & Motion Planning
- Object recognition & Avoidance
- 0.06m2 & **800g of weight**
- Battery Capacity **5410mAh**

Nano-drone

https://www.bitcraze.io/products/crazyflie-2-1



- Smaller form factor of 0.008m2
- Weight **27g (30X lighter)**
- Battery capacity **250mAh (20X smaller)**

Can we fit sufficient intelligence in a 30X smaller payload, 20X lower energy budget?

© 2024 Alberto Sangiovanni Vincentelli
Luca Benini

Achieving True Autonomy on Nano-UAVs

Multiple, complex, heterogeneous tasks at high speed and robustness **fully on board**

Object detection

Obstacle avoidance & Navigation

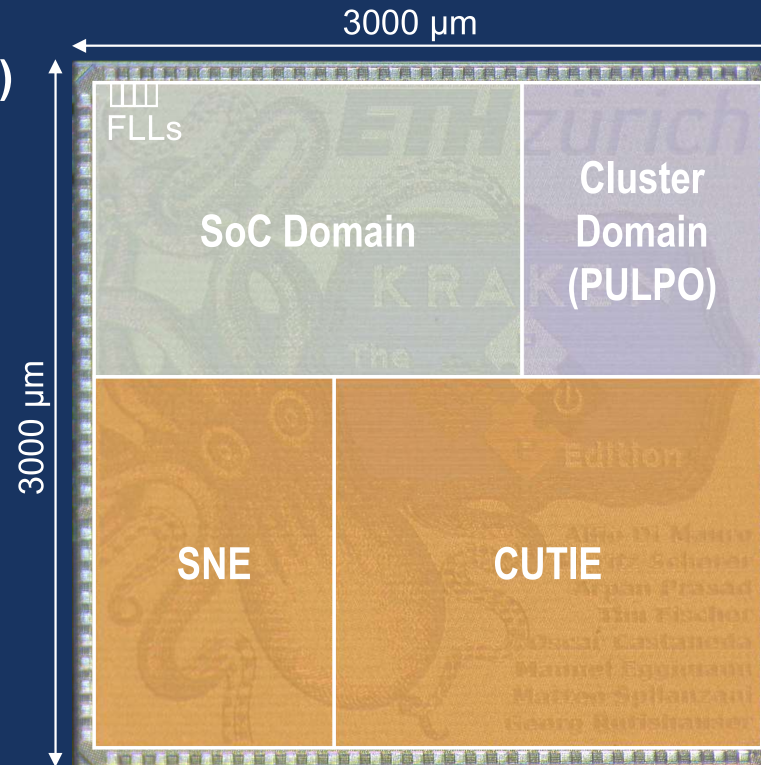
Environment exploration

Multi-GOPS workload at extreme efficiency → P_{max} 100mW

Kraken: 22nm SoC, Multiple Heterogeneous Accelerators

The *Kraken*: an "Extreme Edge" Brain

- **RISC-V Cluster (8 Cores + 1)**
- **CUTIE – dense ternary neural network accelerator**
- **SNE – energy-proportional spiking neural network accelerator**

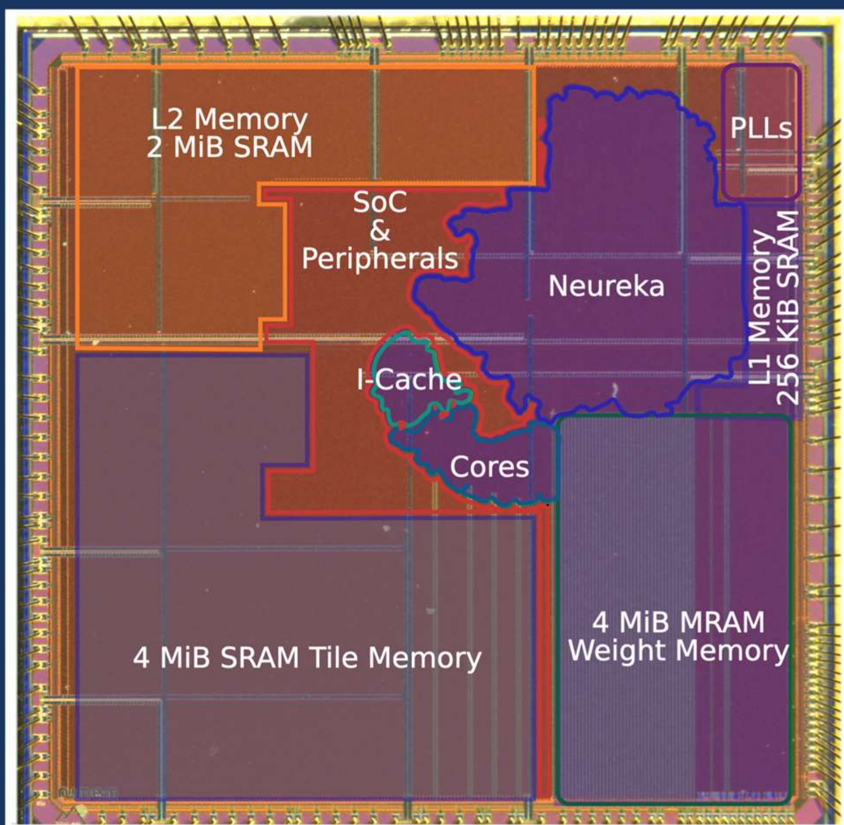


[Di Mauro HotChips22]

Technology	22 nm FDSOI
Chip Area	9 mm^2
SRAM SoC	1 MB
SRAM Cluster	128 KB
VDD range	0.55 V - **0.8 V**
Cluster Freq	**~370MHz**
SNE Freq	**~250MHz**
CUTIE Freq	**~140MHz**

© 2024 Alberto Sangiovanni Vincentelli

Siracusa: 16nm SoC, Tightly Coupled at MRAM Accelerator



	Vega [1]	Diana [2]	Marsellus [3]	[4]	[5]	*Siracusa*
Technology	22nm FDX	22nm FDX	22nm FDX	40nm	22nm	16nm FinFET
Area	10mm²	10.24mm²	8.7mm²	25mm²	8.76mm²	16mm²
On-chip mem	1728 KB SRAM 4 MB MRAM (L3)	896 KB SRAM	1152 KB SRAM	768 KB	1428 KB	6400 KB SRAM 4 MB MRAM (L1)
Peak Perf 8b	32.2 GOPS	140 GOPS	90 GOPS	N/A	146 GOPS	**698 GOPS**
Peak Eff 8b	1.3 TOPS/W	2.07 TOPS/W	1.8 TOPS/W	0.94 TOPS/W	0.7 TOPS/W	**2.68 TOPS/W**
Peak Eff (WxAb)	1.3 TOPS/W	4.1TOPS/W (2x2b) **600 TOPS/W** **(analog)**	12.4 TOPS/W (2x2b)	60.6 TOPS/W (1x1b)	0.7 TOPS/W	8.84 TOPS/W (2x8b)
Area Eff	3.2 GOPS/mm²	21.2 GOPS/mm²	47.4 GOPS/mm²	N/A	58.3 GOPS/mm²	**65.2 GOPS/mm²**

[1] D. Rossi et al., JSSC'21
[2] P. Houshmand et al., JSSC'23
[3] F. Conti et al., JSSC'23
[4] M. Chang et al., ISSCC'22
[5] Q. Zhang et al., VLSI Symposium'22

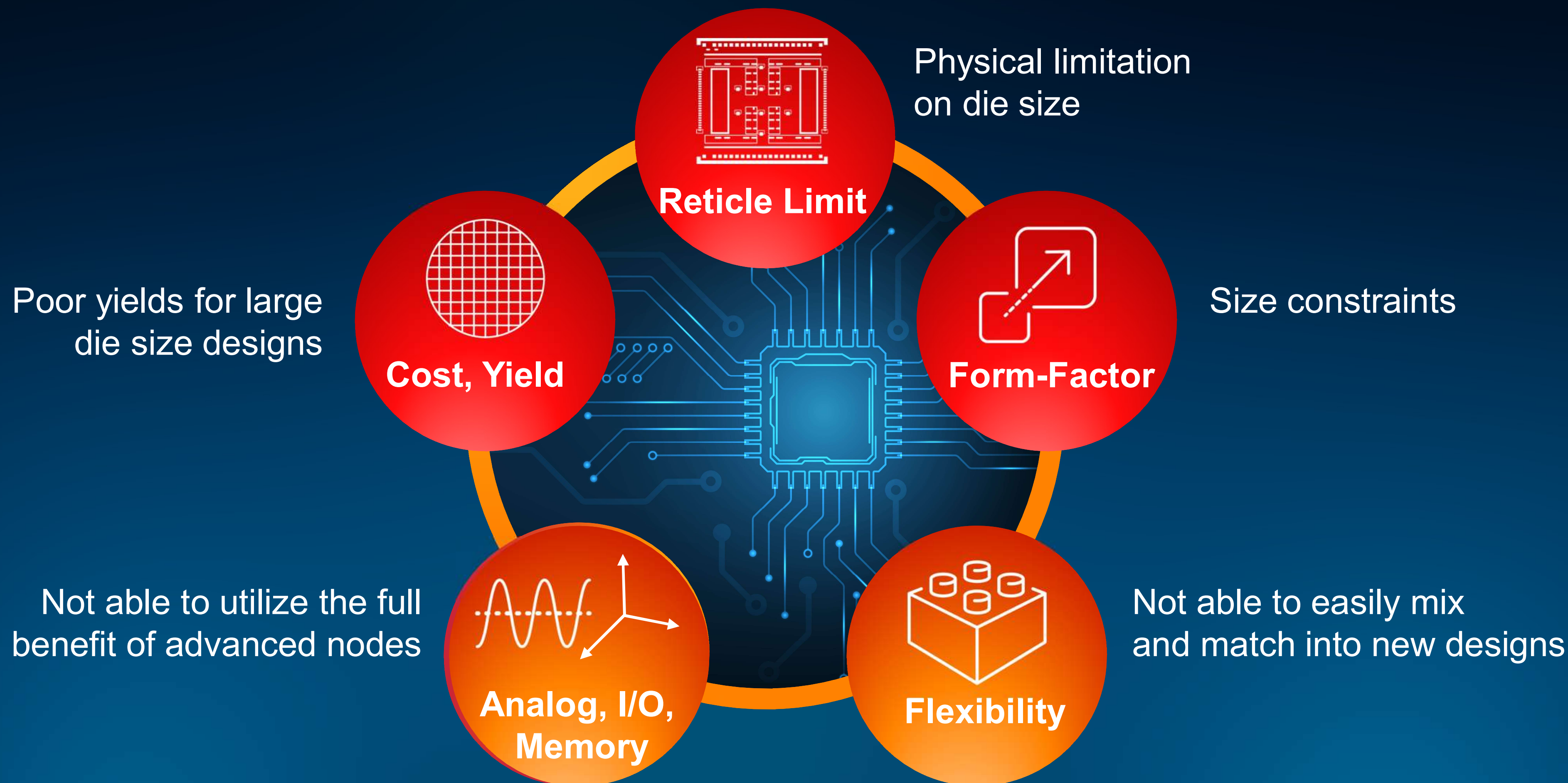
Balance efficiency, peak performance, area efficiency without compromises in precision

N-EUREKA 36-cores configuration

[A. Prasad et al., "Siracusa: a 16nm Heterogeneous RISC-V SoC for Extended Reality with At-MRAM Neural Engine," IEEE Journal of Solid-State Circuits (accepted)]

ETH zürich © 2024 Alberto Sangiovanni Vincentelli

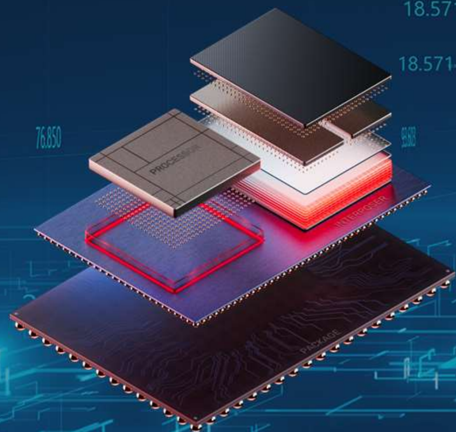
Monolithic 2D-IC Design Limitations

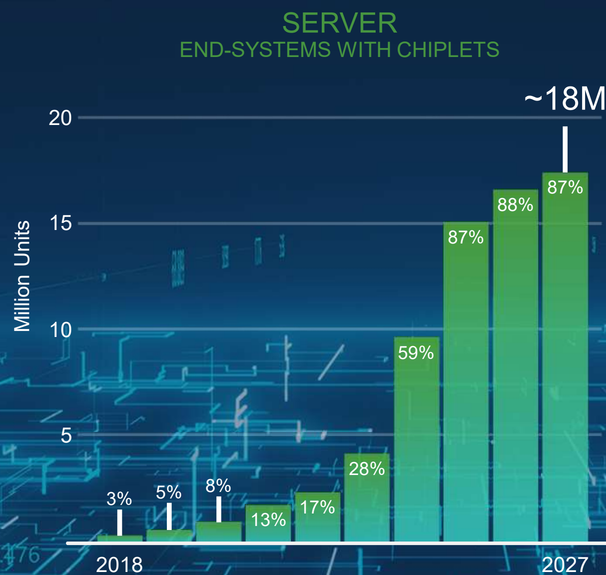


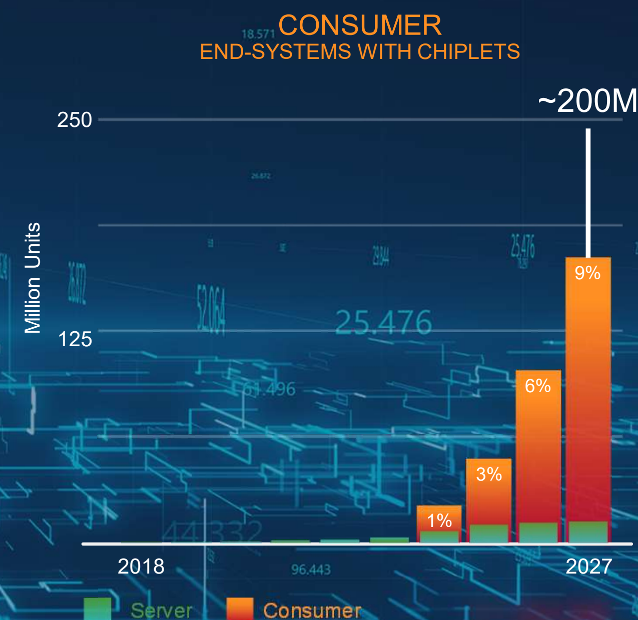
© 2024 Alberto Sangiovanni Vincentelli

AI Driving 3D-IC Proliferation

Productivity and cost improvements are critical







© 2024 Alberto Sangiovanni Vincentelli

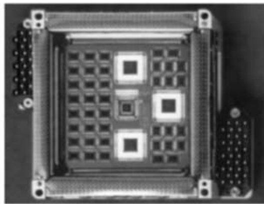
Berkeley UNIVERSITY OF CALIFORNIA cadence

Outline

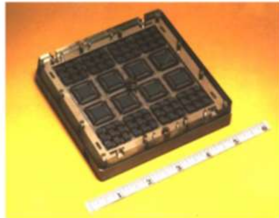
- Big chips or chiplet-based implementation?
- **Early Chiplet Examples**
- The Design Problem
- Is AI a Panacea?
- Design Flows for Chiplet-based design
- A Bit of Research

© 2024 Alberto Sangiovanni Vincentelli

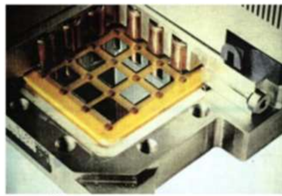
Déjà vu all over again?



Source: DEC



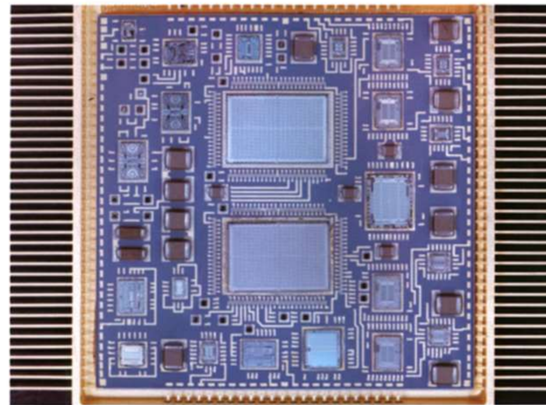
Source: Motorola/Siemens



Source: IBM

Source: ICE, "Roadmaps of Packaging Technology" 22555

Figure 12-1. Early MCMs in ECL Mainframe Computers in the Mid 1980's



Source: Boeing Microelectronics/ICE, "Roadmaps of Packaging Technology" 22563

Figure 12-12. Rocket Control and Monitoring Hybrid

Courtesy of nChip/ICE, "Roadmaps of Packaging Technology" 22033

Figure 12-68. Ross hyperSPARC™ Module With Wirebonded Die on a Chip Substrate

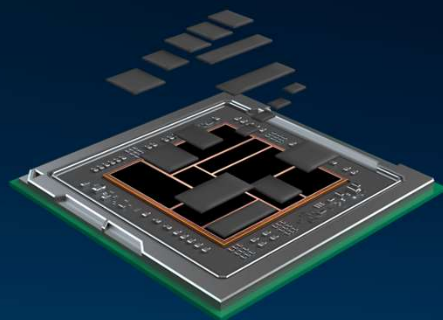
MCM Start-ups in the 80s:

Polycon, Advanced Packaging Systems, ISA, Polylithics, Alcoa Microelectronics, nChip, TIÕs HDI and Pacific Microelectronics Center

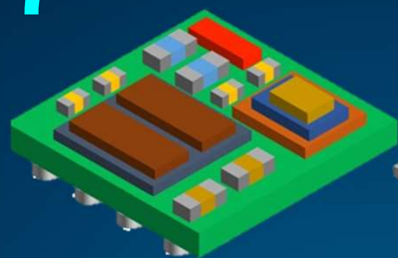
© 2024 Alberto Sangiovanni Vincentelli

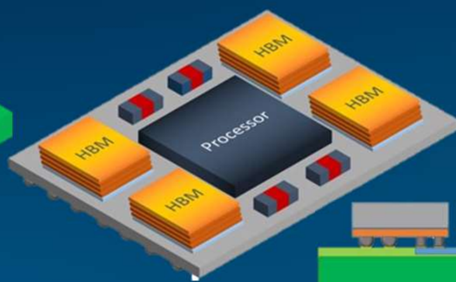
Berkeley UNIVERSITY OF CALIFORNIA cadence

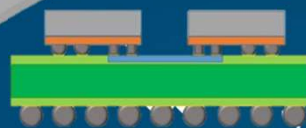
Heterogenous Integration: Multiple Packaging Technologies

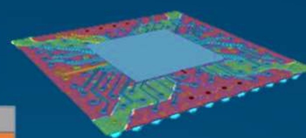


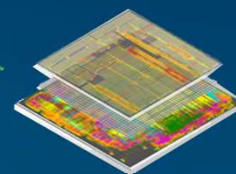
Heterogeneous Integration

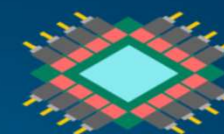










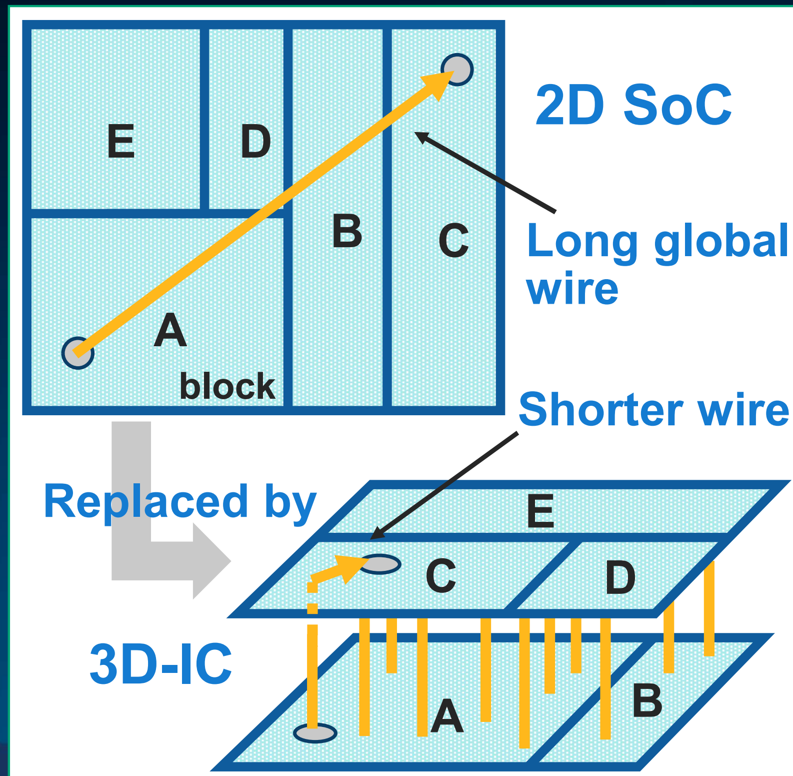


System in Package (SiP/MCM)	2.5D-IC (Silicon/RDL Interposer)	Interconnect Bridges	Ultra-High-Density RDL (FOWLP)	Silicon Stacking	3D System-on-a-Wafer	Co-Packaged Optics
1990	2010	2012	2015	2018	2020	2022

© 2023 Alberto Sangiovanni Vincentelli

Berkeley UNIVERSITY OF CALIFORNIA
cadence

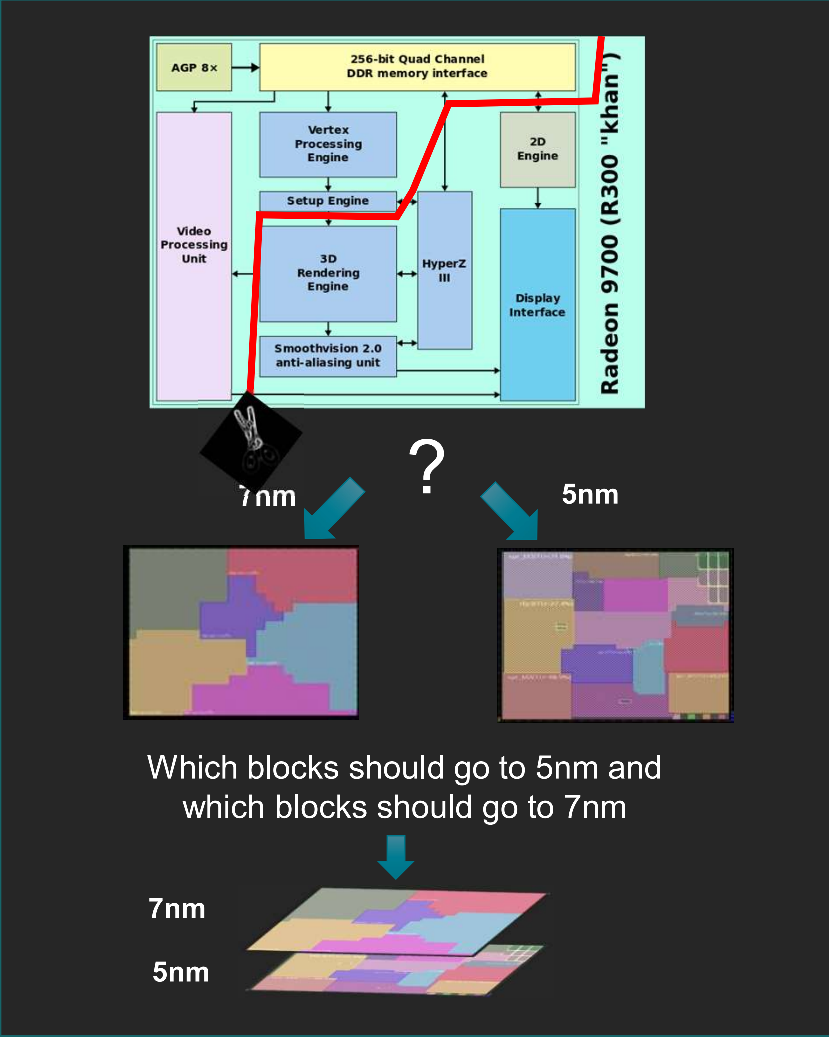
Silicon Stacking (3D-IC)

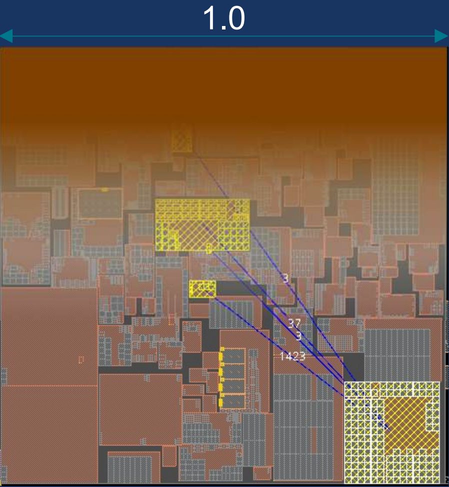


- Shorter Wire
- Less Power
- Higher Performance
- Higher Bandwidth
- Smaller Profile
- Better Yield

© 2024 Alberto Sangiovanni Vincentelli

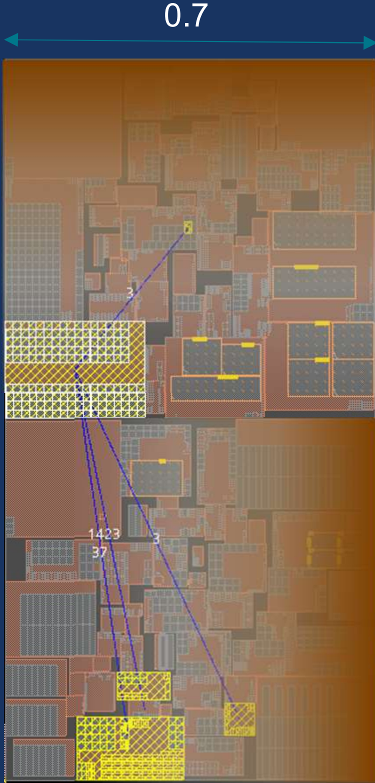
Module-Level 3D-IC Partitioning





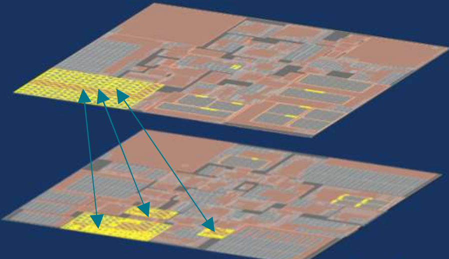
2D module placement

vs.



Shorter wire length in 3D

- Multi-die partitioning and hierarchical floorplan synthesis in one unified algorithm (EFS)
- Consider wire length, bump interface penalty, different process nodes and module utilization on the functional module level

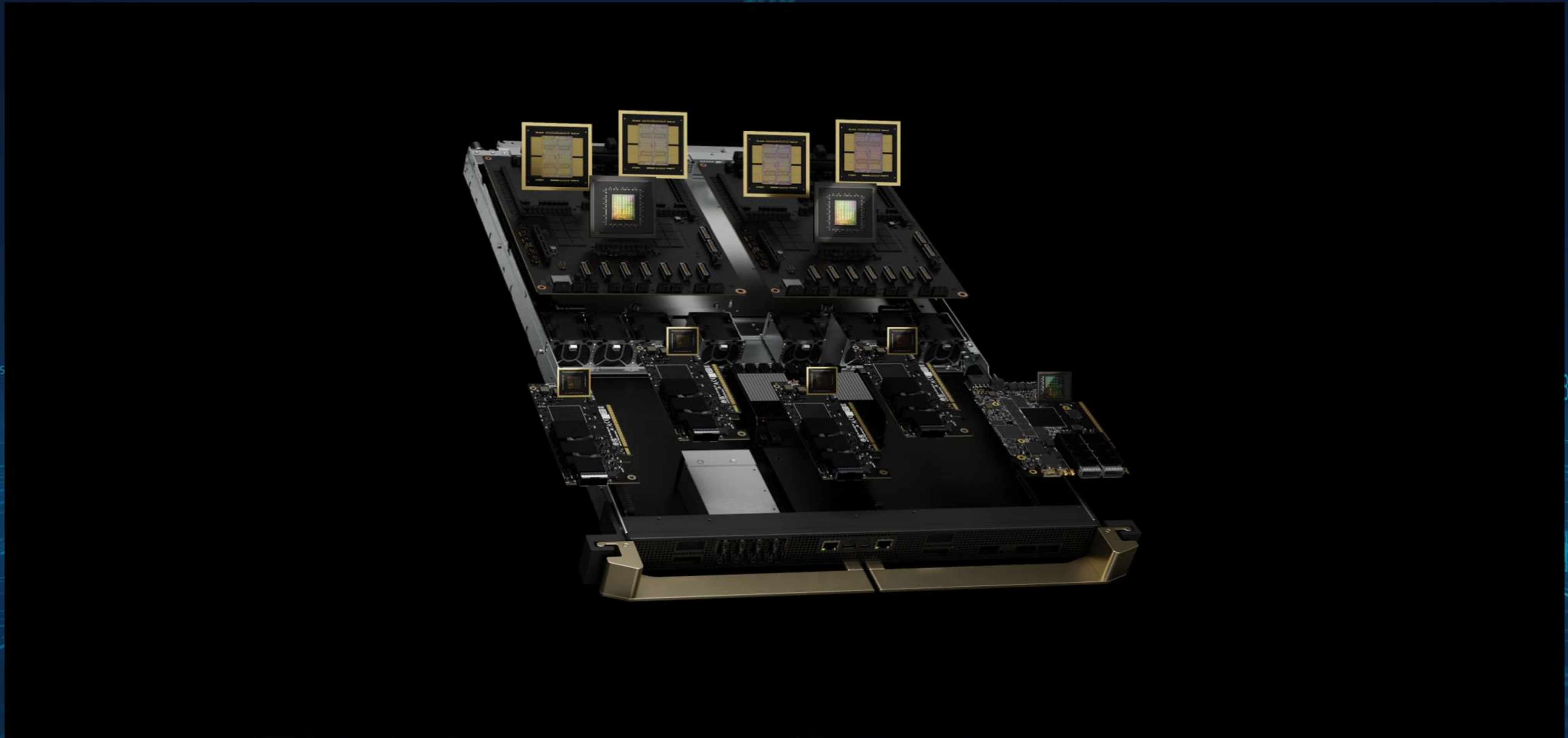


© 2024 Alberto Sangiovanni Vincentelli
Berkeley UNIVERSITY OF CALIFORNIA cadence

Nvidia Blackwell Superchip

© 2024 Alberto Sangiovanni Vincentelli

Nvidia Blackwell Compute Tray



© 2024 Alberto Sangiovanni Vincentelli

Marvell MoChi Architecture **(JSSCC '15 Key note: Sehat Sutardja)**

MoChi – Mix & Match

Off Package Example

Platform A

Platform B

Armada Chip

- Optimized solution per platform
- No unnecessary cost burdens
- Daisy chain for even more options

Eth. Switch

Eth

SATA

USB3.0

Marvell SERDES Lanes

Armada Chip

Company Portfolio

MARVELL

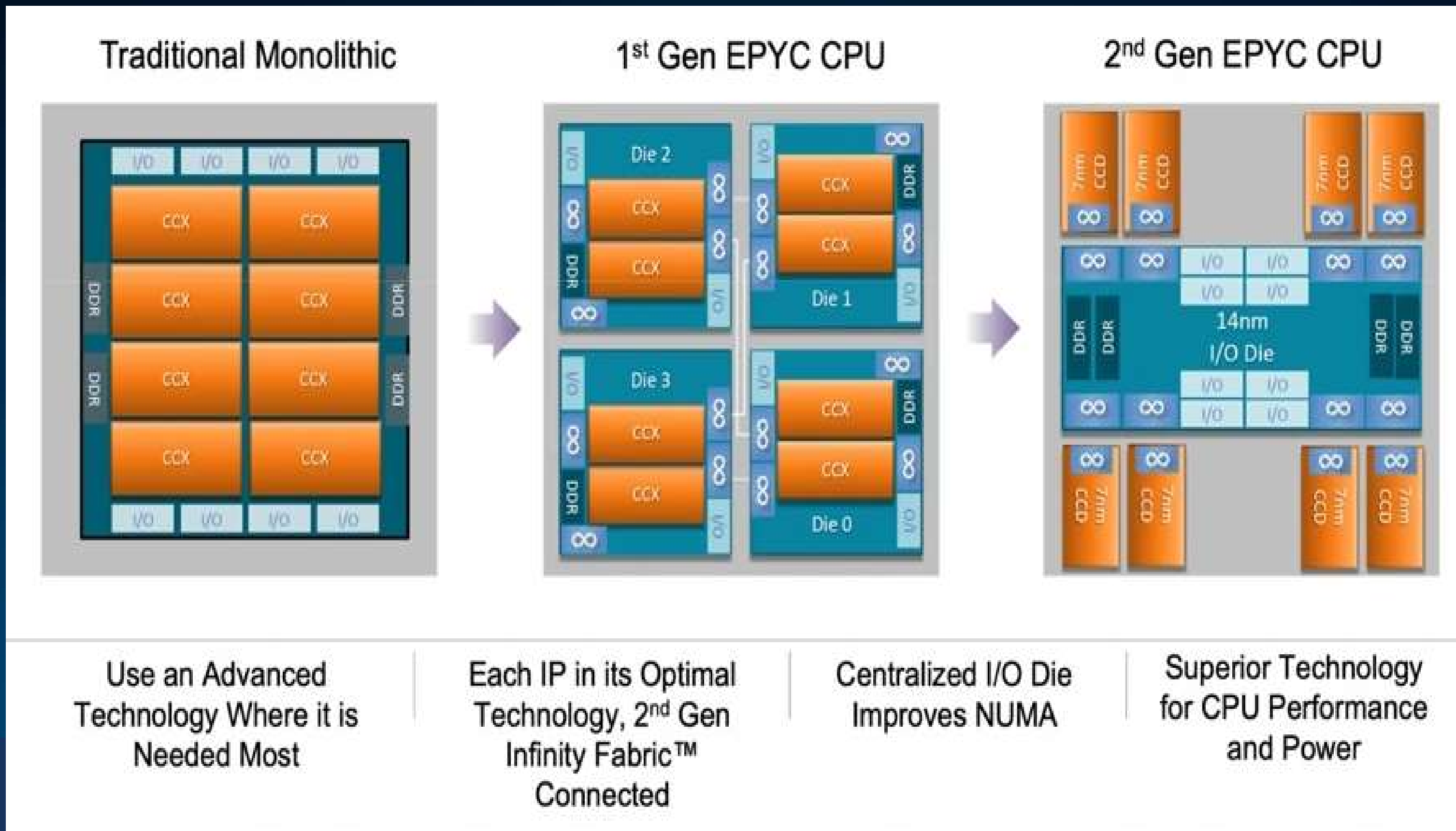
© 2015 Marvell Confidential. All rights reserved.

JSSCC Key note: Sehat Sutardja

Marvell Company Newsroom 2015

© 2024 Alberto Sangiovanni Vincentelli

AMD Zen Chiplet Architecture



1st gen: 10% additional silicon real estate for

- die-to-die communication blocks,
- redundant logic
- other unnamed add-ons

BUT 41% LOWER COST!

ISSCC, 2021

© 2024 Alberto Sangiovanni Vincentelli

Berkeley UNIVERSITY OF CALIFORNIA

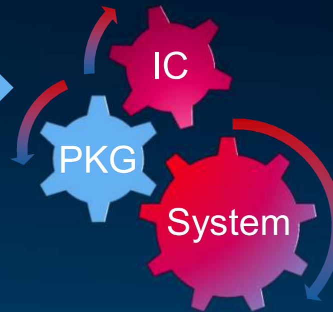
Convergence?

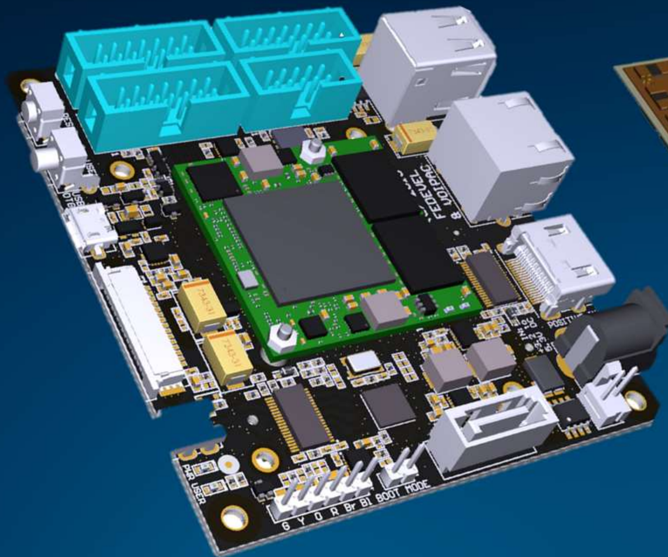
OSATs (SWaP)

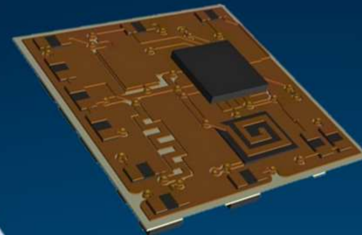
1980-2010

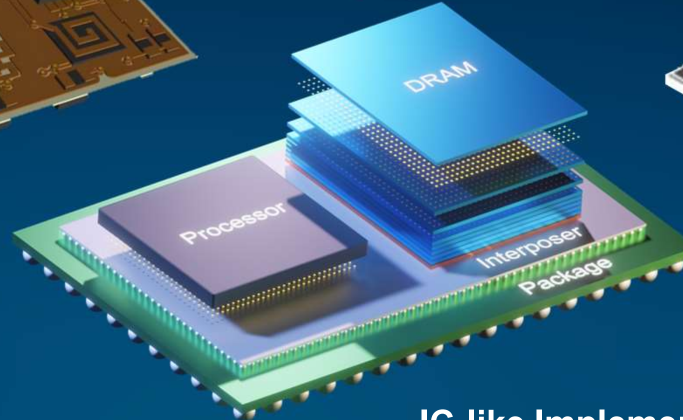
Foundries (PPA)

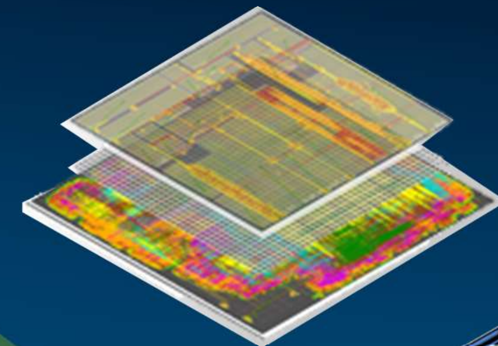
2011-Now

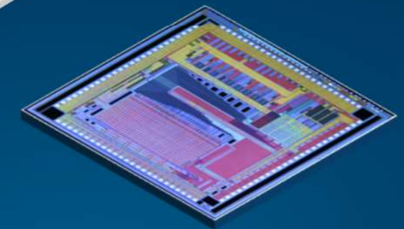












PCB Layout Flow
System-Level Analysis

PCB-like Flow
System-Level Analysis

IC-like Implementation Flow
IC signoff Methodology and
System-Level Analysis

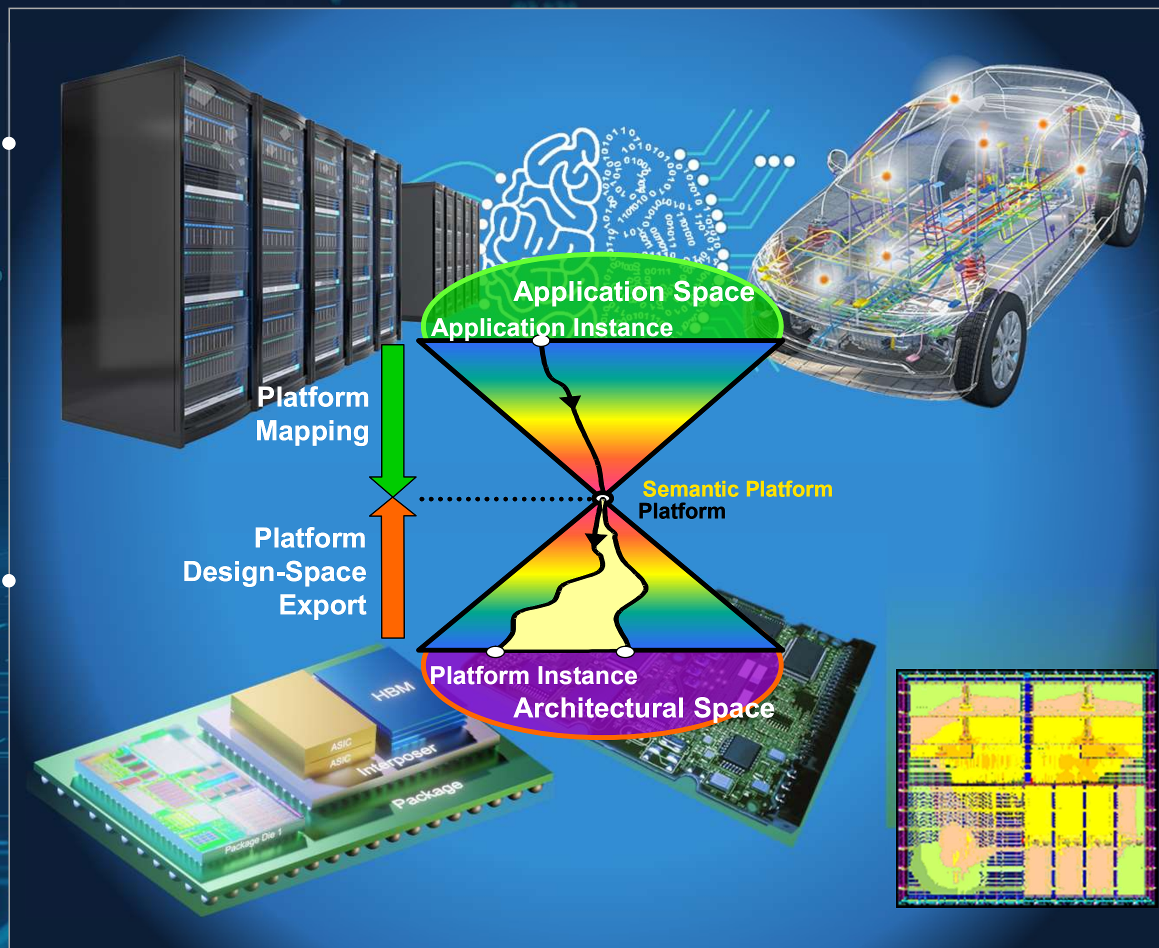
IC Flow
IC signoff Methodology

© 2023 Alberto Sangiovanni Vincentelli

The Playfield: Cyber Physical System Design

Electronics and Software Design

Multiphysics Design



3D-IC System-in-Package Design

Advanced-Node Chip Design

© 2024 Alberto Sangiovanni Vincentelli

Berkeley UNIVERSITY OF CALIFORNIA cadence

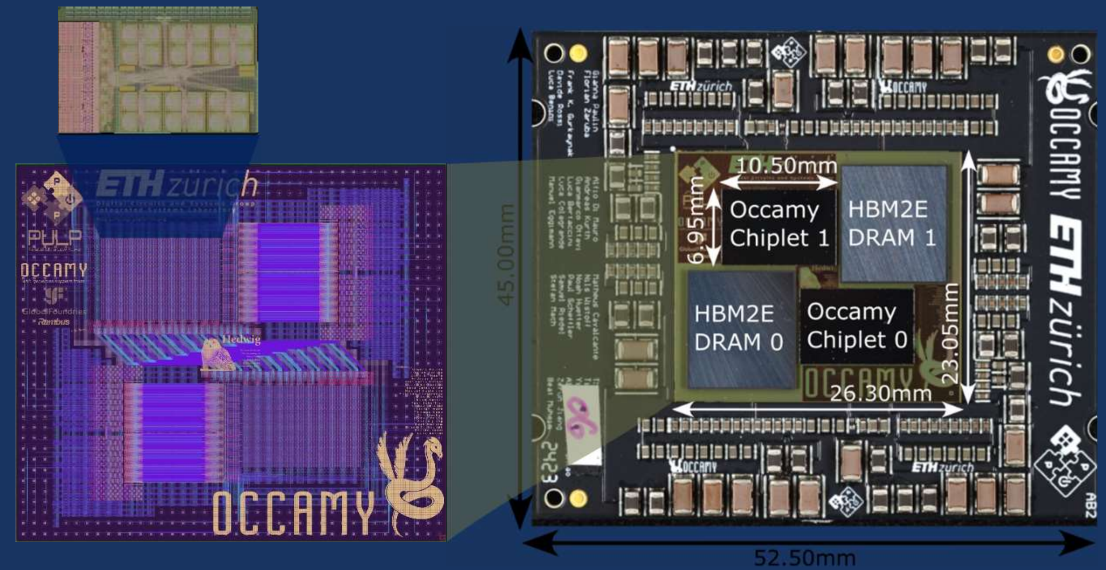
Outline

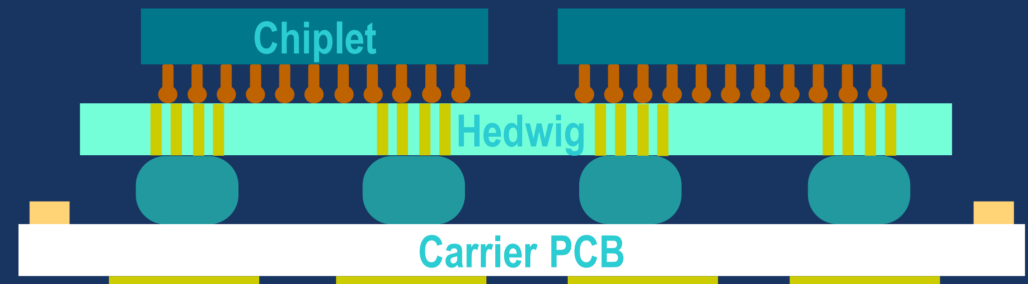
- Big chips or chiplet-based implementation?
- Early Chiplet Examples
- **The Design Problem**
- Is AI a Panacea?
- Design Flows for Chiplet-based design
- A Bit of Research

© 2024 Alberto Sangiovanni Vincentelli

Benini: Occamy 2.5D System: Chiplets on Passive Interposer

- *Hedwig* interposer
  - 65nm, passive (BEOL only)
  - Connects 2× 73 mm² Occamy chiplet (GF 12LP+) and 2× Micron HBM2E
  - Distributes power and IOs
- Carrier PCB
  - RO4350B (low CTE, high stability)
  - LGA 2011 pinout adapted to fit ZIF socket
  - Stabilizes assembly and power
- Occamy system module
  - 432+2 RISC-V cores, 32 GiB HBM2E
  - 768 DP-GFLOP/s peak performance (HPC)
  - 6144 FP8-GFLOP/s peak performance (ML)

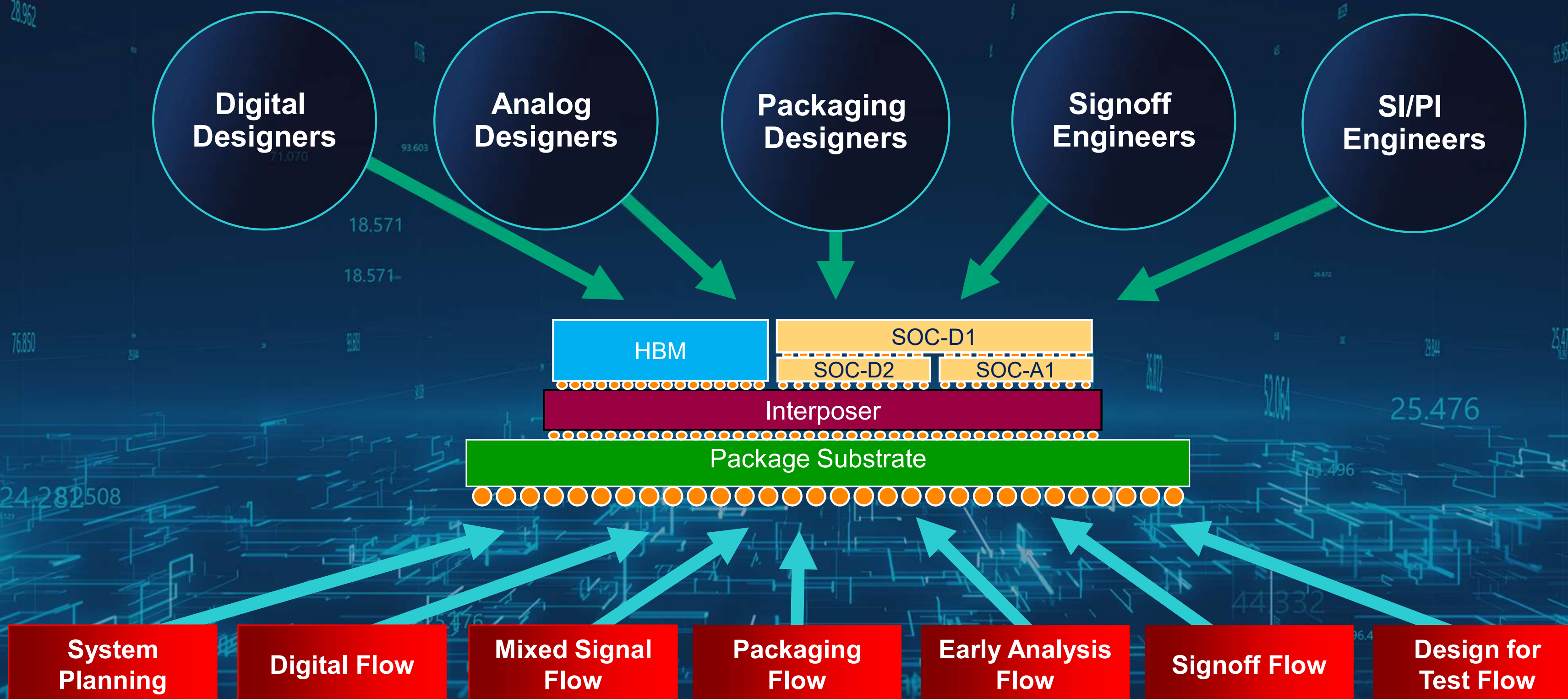




ETH zürich © 2024 Alberto Sangiovanni Vincentelli

Chiplet 3D-IC Design Flow – Different Requirements based on users

Requires a Platform for full solution yet modular for different teams



3D-IC Flow is a multi-engineering discipline methodology
Requires tight collaboration across all teams.
The cross-over is between SoC, Chiplet and Packaging Design Methodology

© 2024 Alberto Sangiovanni Vincentelli

Huge design space: Exploration problem

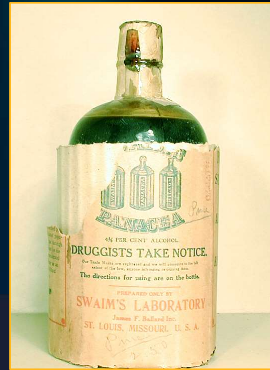
© 2024 Alberto Sangiovanni Vincentelli

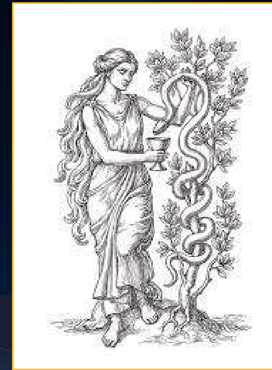
Outline

- Big chips or chiplet-based implementation?
- Early Chiplet Examples
- The Design Problem
- **Is AI a Panacea?**
- Design Flows for Chiplet-based design
- A Bit of Research

© 2024 Alberto Sangiovanni Vincentelli

Is AI a Panacea?





171 Talks	**6** Keynotes and visionary talks
1 Workshop	**2** Panels
6 Tutorials	**24** Sessions

AI
ARTIFICIAL INTELLIGENCE

@2023 DAC

© 2024 Alberto Sangiovanni Vincentelli

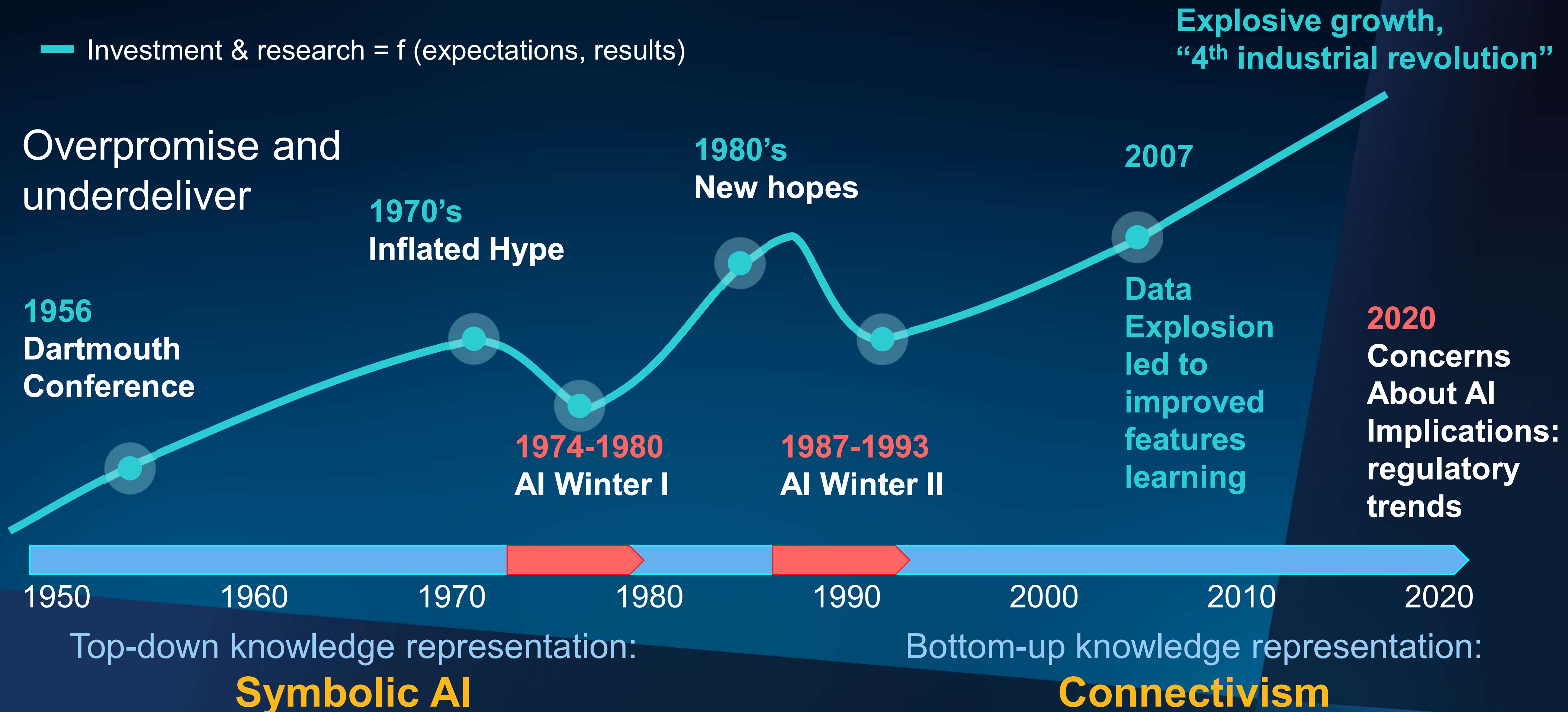
Definitions

Artificial Intelligence: The theory and development of computer systems able to perform tasks normally requiring human intelligence (Oxford Dictionary)

Machine Learning: algorithms and supporting theory for making predictions and decisions **under uncertainty** based on **observed data.**

© 2024 Alberto Sangiovanni Vincentelli
Berkeley UNIVERSITY OF CALIFORNIA cādence

Winters of AI

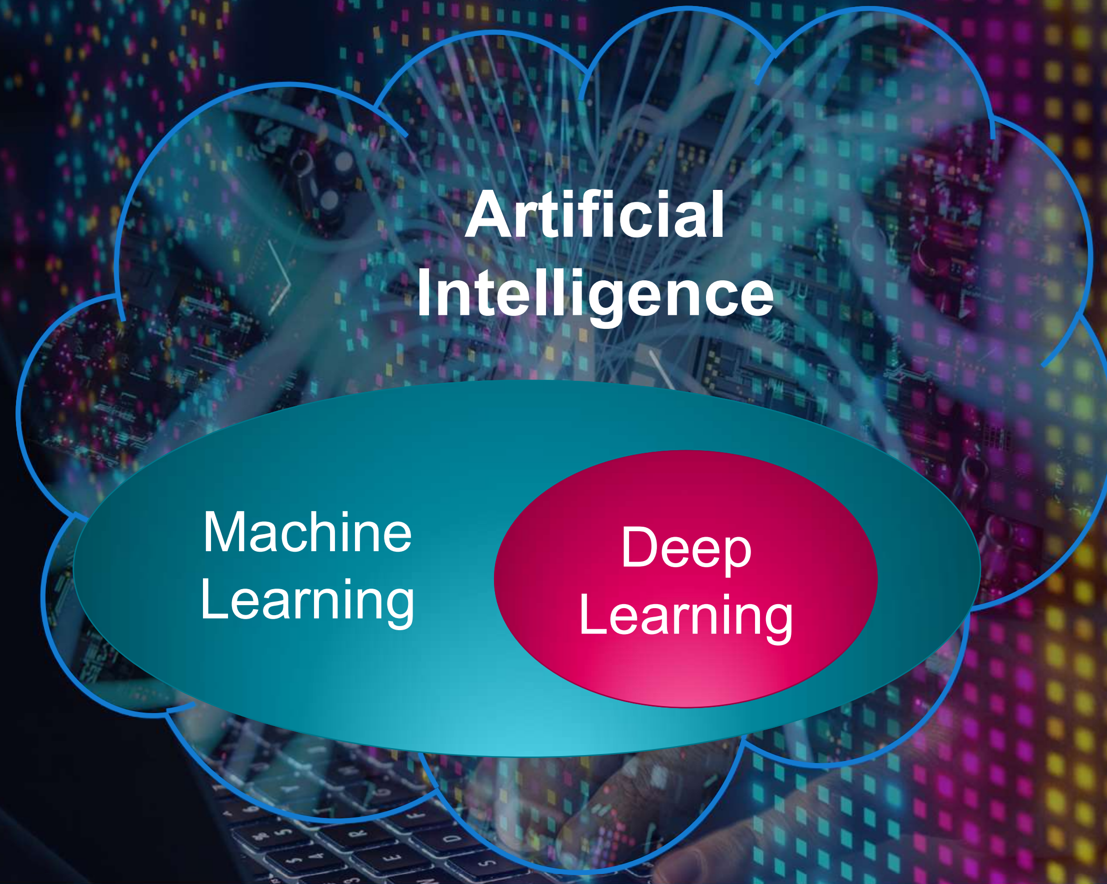


Credit: Clive Thompson, Medium

© 2024 Alberto Sangiovanni Vincentelli

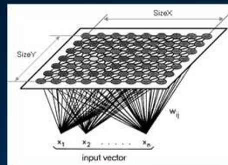
Berkeley UNIVERSITY OF CALIFORNIA

AI/Machine Learning/Deep Learning

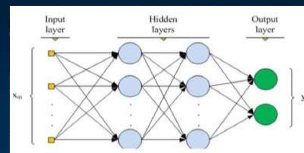


© 2023 Alberto Sangiovanni Vincentelli

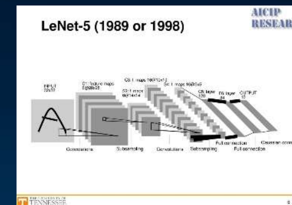
Machine Learning



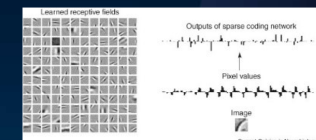
Hopfield network, SOM (Kohonen, 1982), Neural PCA (Oja, 1982)



Multilayer perceptrons and backpropagation (Rumelhart et al., 1986)

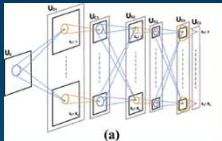


Autoencoders (Baldi&Hornik,1989)
Convolutional network (LeCun, 1989)

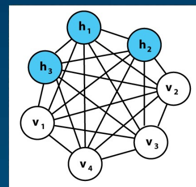


Sparse coding (Field, 1993)

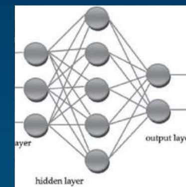
Rosenblatt's Perceptron



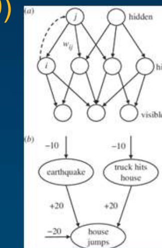
Neocognitron (Fukushima, 1980)



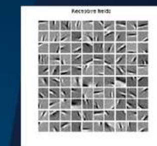
Boltzmann machines (Ackley et al,.1986)



RBF networks (Broomhead &Lowe, 1988)



Sigmoid belief network (Neal, 1992)



(Olshausen, 1996)

2000s Sparse, Probabilistic, and Energy models (Hinton, Bengio, LeCun, Ng)

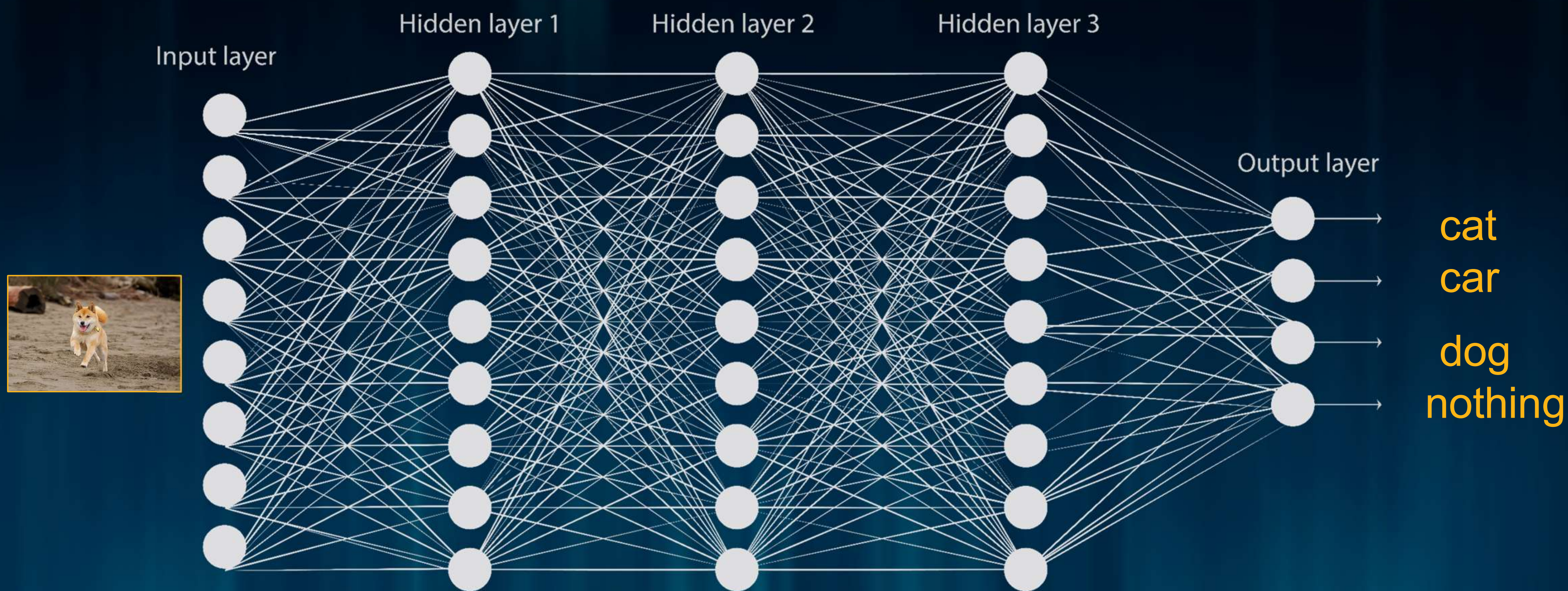
1958 1980 1982 1985 1986 1988 1989 1992 1993 2000

Is deep learning 3, 30, or 60 years old?

based on history by K. Cho

© 2024 Alberto Sangiovanni Vincentelli
Berkeley UNIVERSITY OF CALIFORNIA

Deep Learning: Many Layer Neural Network



Neural Net Training: Find the weights that minimize the difference between labels and activation.

© 2024 Alberto Sangiovanni Vincentelli

Machine Learning and Approximations

A deep learning architecture is a multilayer function with many parameters

Parameters are determined by fitting a training set and verified using a test set

Is there any guarantee that this function will approximate the «real» function?

© 2024 Alberto Sangiovanni Vincentelli

Generative AI

- Generative artificial intelligence (generative AI, GenAI, or GAI) is artificial intelligence capable of **generating text, images, videos, or other data** using generative models, often in response to prompts. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics.

- A generative model is a statistical model of the joint probability distribution $P(X,Y)$ on a given observable variable X and target variable Y;

  A generative model can be used to "generate" random instances (outcomes) of an observation x

Transformers

- Text is converted to **numerical representations** called **tokens**, and each token is converted into a vector via looking up from a word embedding table. At each layer, each token is then contextualized within the scope of the context window with other (unmasked) tokens via a parallel multi-head attention mechanism allowing the signal for key tokens to be amplified and less important tokens to be diminished.

Large Language Models (LLMs)

- A **large language model (LLM**) is a language model to achieve general-purpose language understanding and generation. LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process. LLMs are artificial neural networks. The largest and most capable, as of March 2024, are built with a decoder-only transformer-based architecture.

- LLMs include ChatGPT 4o with 1 Trillion parameters, Beijing Academy of Artificial Intelligence's Wu Dao 2.0, with 1.75 trillion parameters; Google's Switch Transformer, with 1.6 trillion parameters; Microsoft and Nvidia's MT-NLG, with 540 billion parameters; Hugging Face's Bloom, with 176 billion parameters; Google's LaMDA, with 137 billion parameters, Meta LLAMA 2 70Billion parameters.

Outline

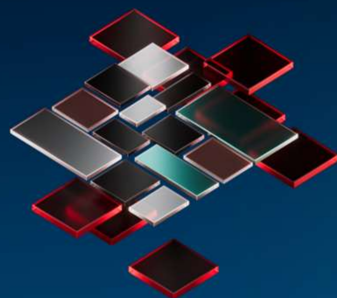
- Big chips or chiplet-based implementation?
- Early Chiplet Examples
- The Design Problem
- Is AI a Panacea?
- **Design Flows for Chiplet-based design**
- A Bit of Research

© 2024 Alberto Sangiovanni Vincentelli

Where and How to Use ML in IC Design?

© 2023 Alberto Sangiovanni Vincentelli

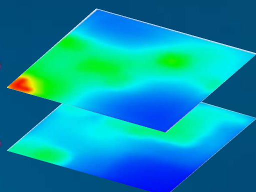
Accelerating 3D-IC and Chiplet Designs



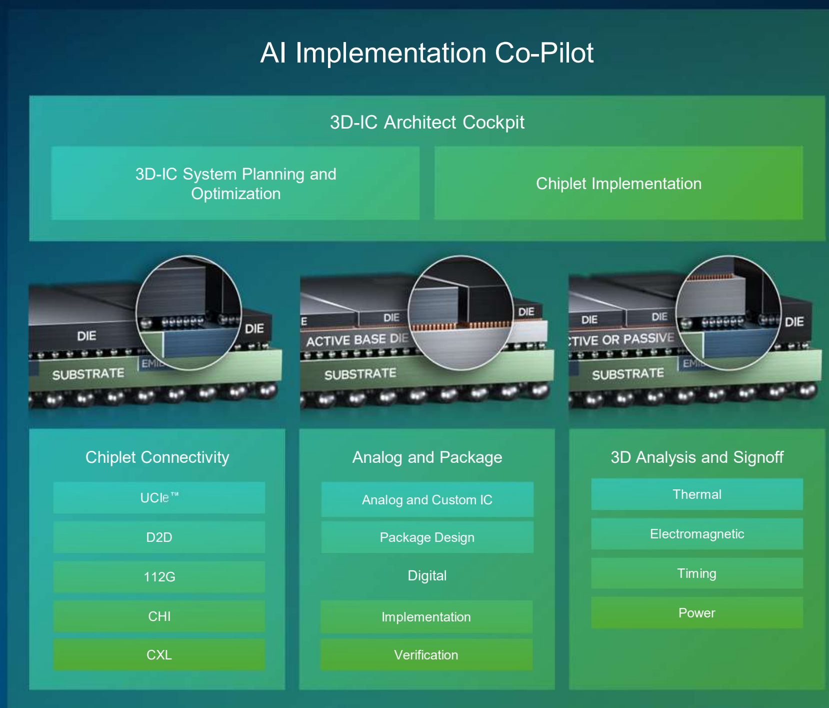
Architecture and IP Designs

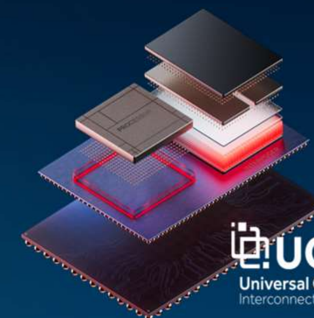
TOP DIE 72.5°C

BOTTOM DIE 65.7°C



Integrity™ 3D-IC Platform

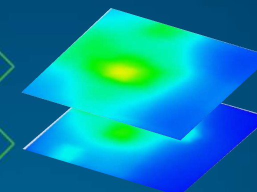




Optimized 3D-IC and Package Designs

TOP DIE 61°C

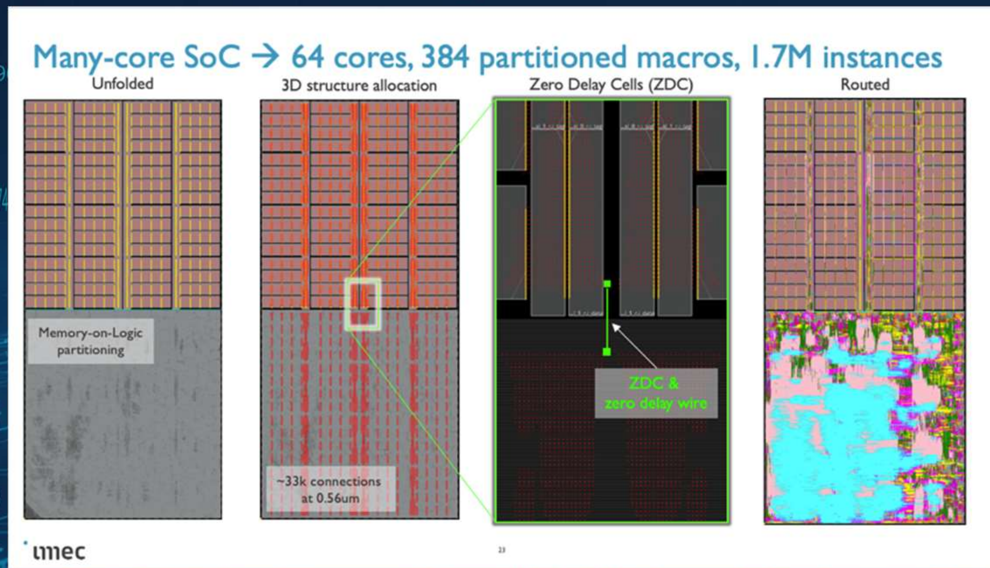
BOTTOM DIE 60°C

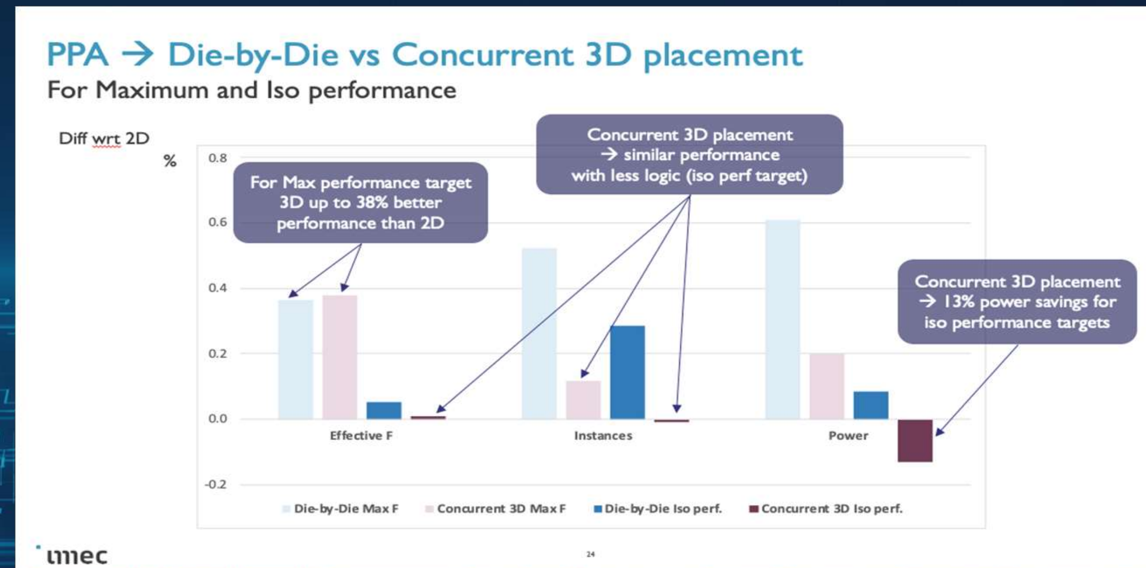


© 2024 Alberto Sangiovanni Vincentelli

Case Study : Imec Experience

"With 3D-IC design continuing to gain momentum, there is an increased need to automate the planning and partitioning of a 3D stack die system more efficiently. "
Eric Beyne, s*enior fellow and program director, 3D System Integration, imec*

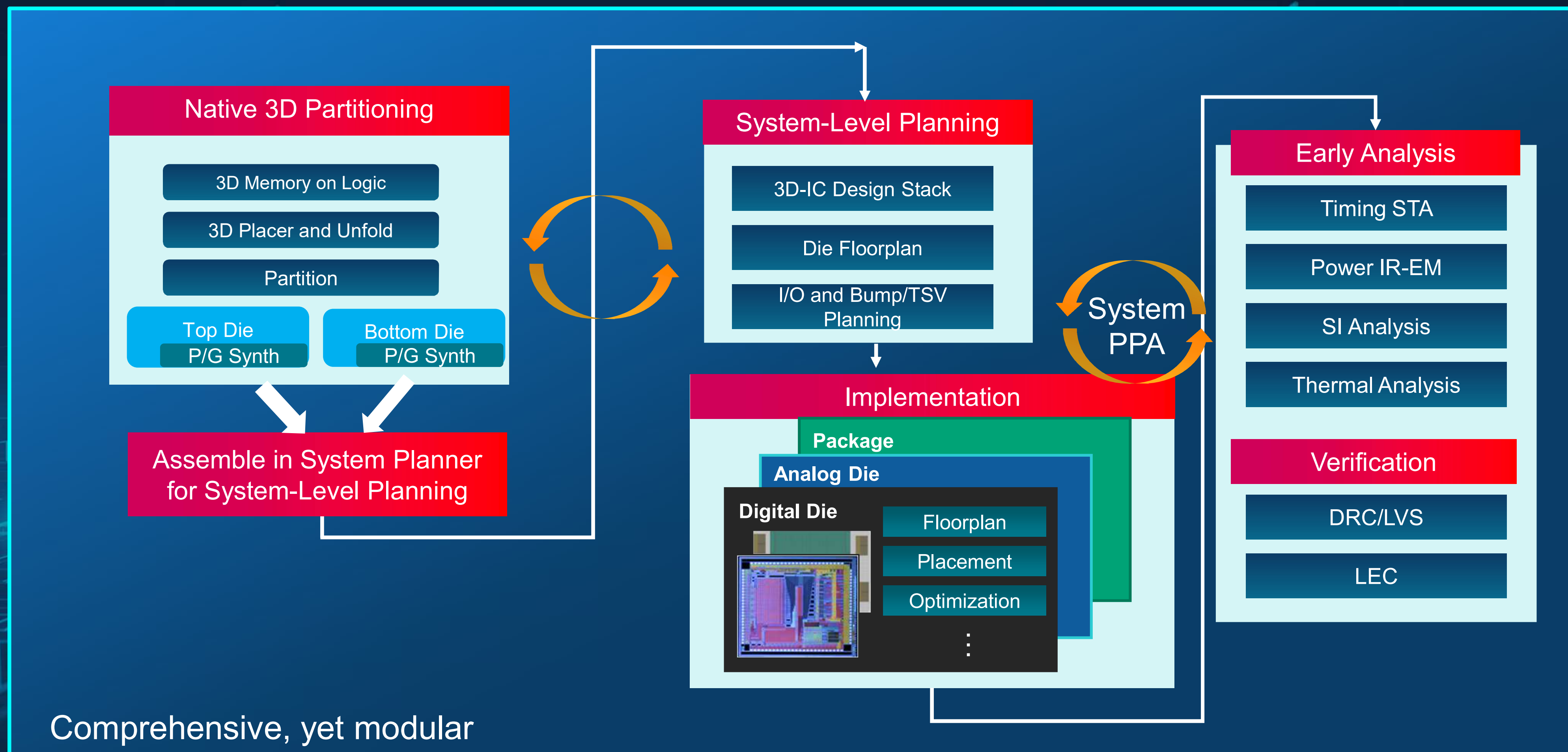




Source : CadenceLive Europe 2021

© 2024 Alberto Sangiovanni Vincentelli

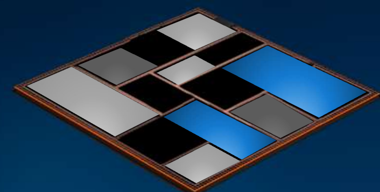
System-Driven PPA



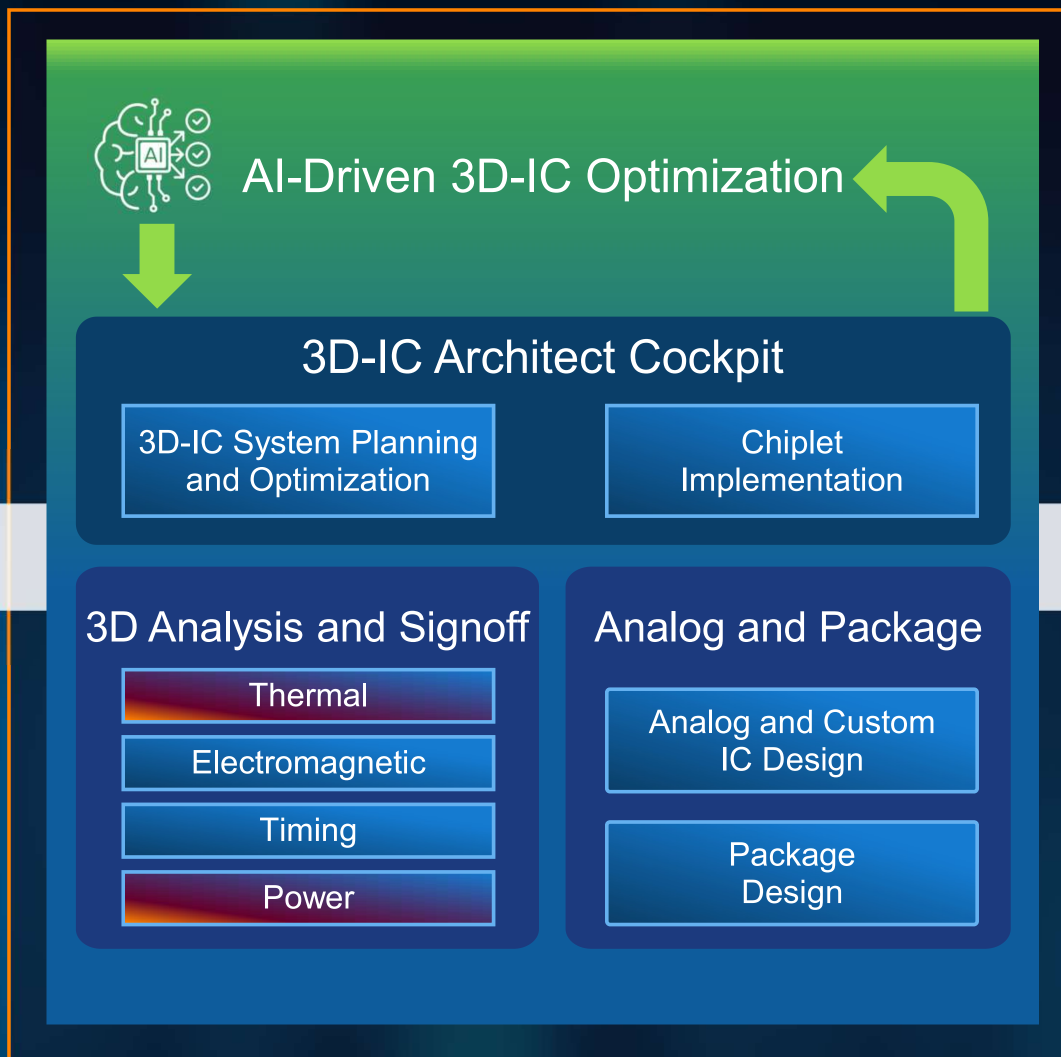
© 2024 Alberto Sangiovanni Vincentelli

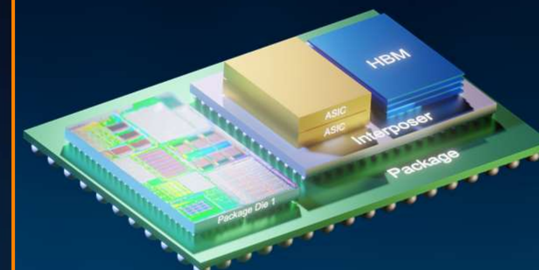
Berkeley UNIVERSITY OF CALIFORNIA cadence

3D-IC Design and Optimization



Design Architecture and IP





Optimized Chiplets and Package Designs

© 2024 Alberto Sangiovanni Vincentelli

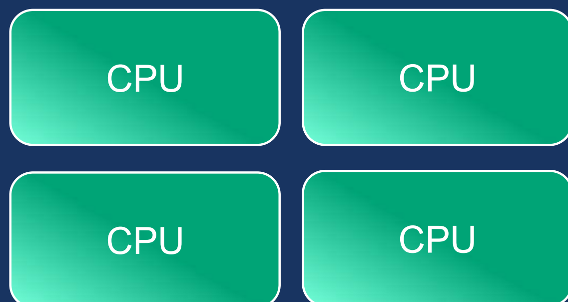
Outline

- Big chips or chiplet-based implementation?
- Early Chiplet Examples
- The Design Problem
- Is AI a Panacea?
- Design Flows for Chiplet-based design
- **A Bit of Research**

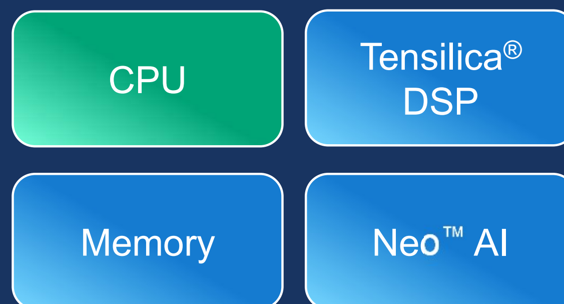
© 2024 Alberto Sangiovanni Vincentelli

Architecting with Chiplets

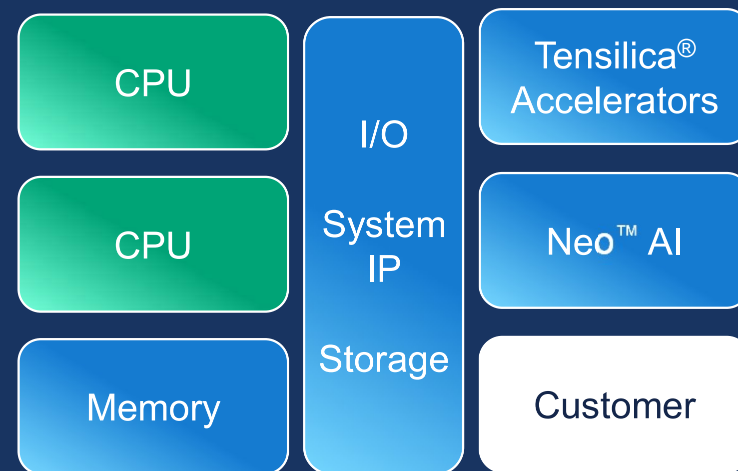
Chiplet "LEGOs" enable *creativity*



Compute



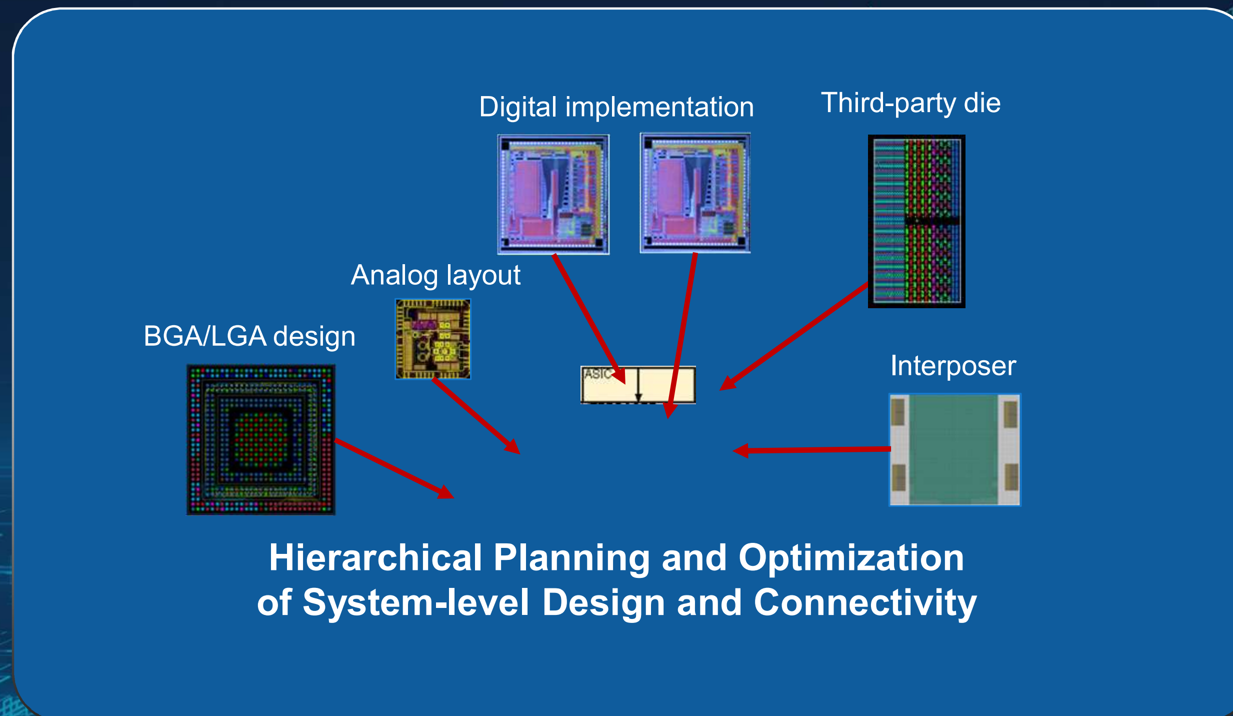
Acceleration



System Hub

© 2024 Alberto Sangiovanni Vincentelli

Goal of Berkeley research

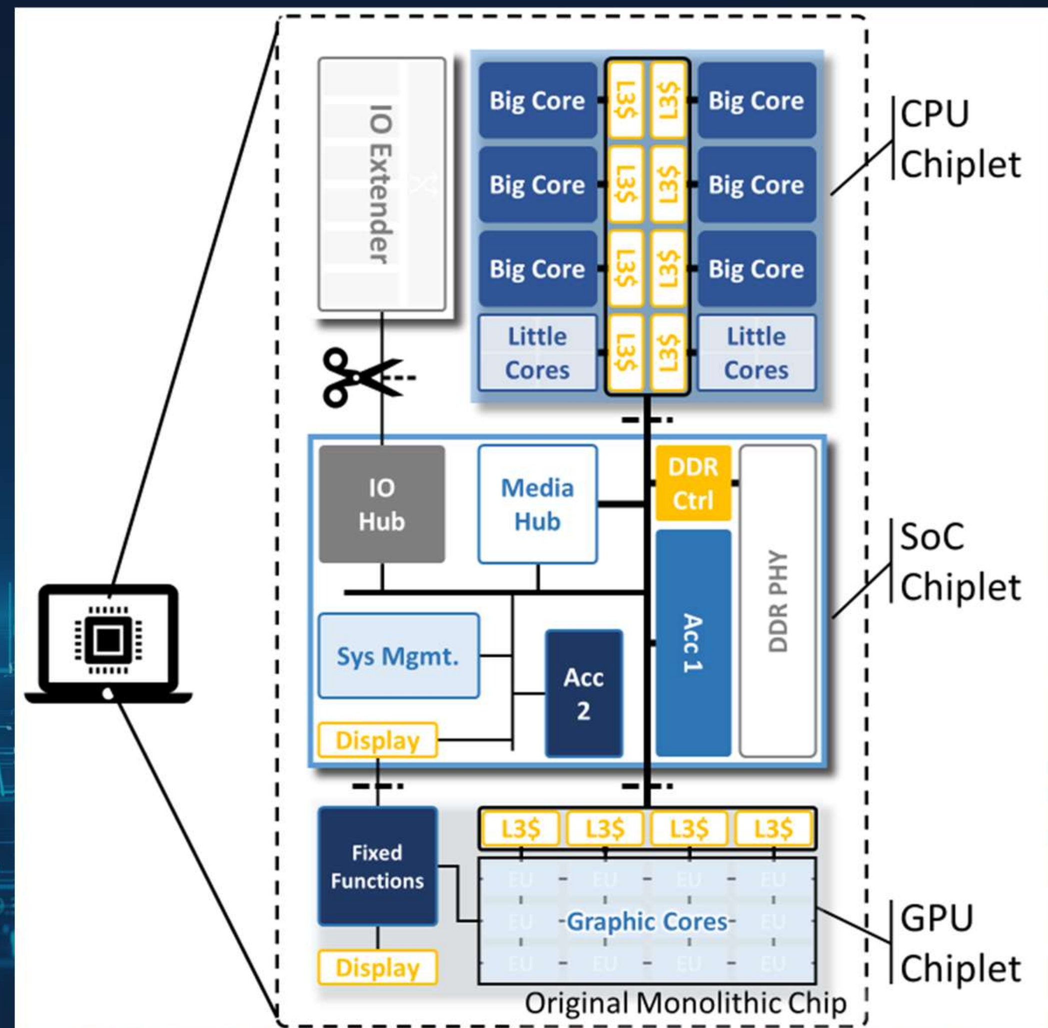


Identifying the partition of original design into components to be implemented as chiplets for optimal performance/cost/re-use potential

© 2024 Alberto Sangiovanni Vincentelli
Berkeley UNIVERSITY OF CALIFORNIA cadence

Re-use paradigm to allow multi-party integration and democratization

- Example
  - Intel Gen 14
  - 4 chiplets for laptops
  - big.LITTLE CPU chiplet

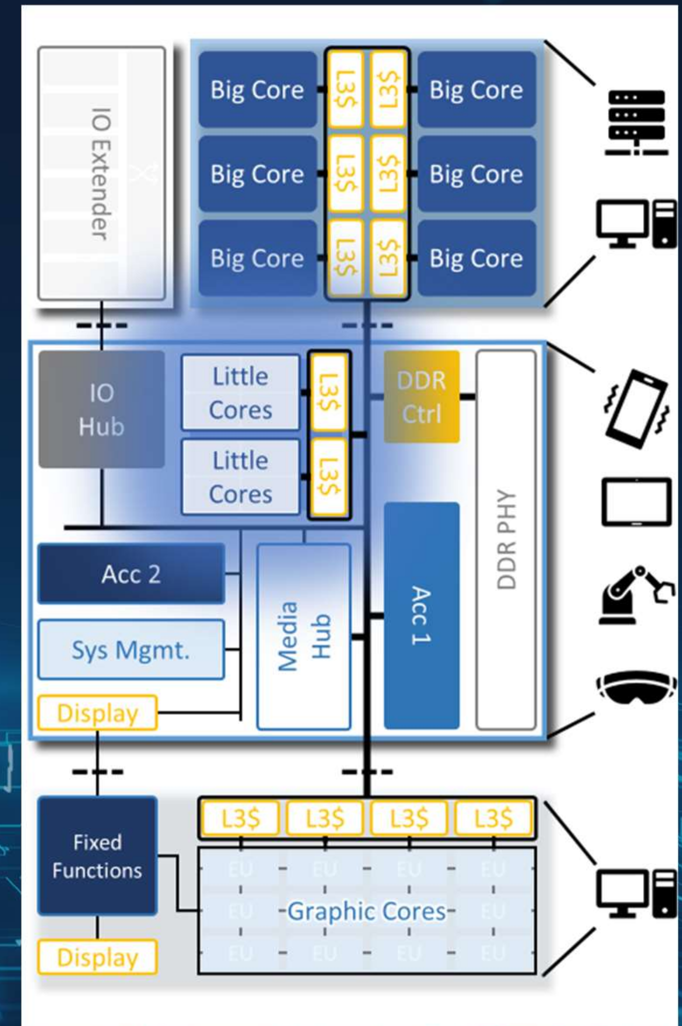


© 2024 Alberto Sangiovanni Vincentelli

Berkeley UNIVERSITY OF CALIFORNIA cadence

Once over lightly

- Reuse chiplets
  - Move small cores to the SoC chiplet
  - Reused in more products
    NRE CO$T --



© 2024 Alberto Sangiovanni Vincentelli

Berkeley UNIVERSITY OF CALIFORNIA cadence

How to Optimize Chiplet Choice?

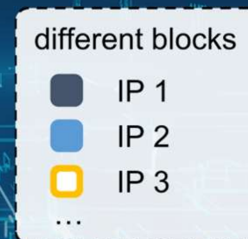
- Designing chiplets is the tradeoff between PPA and BOM
- Designing common chiplet is the tradeoff between PPA, BOM, and NRE

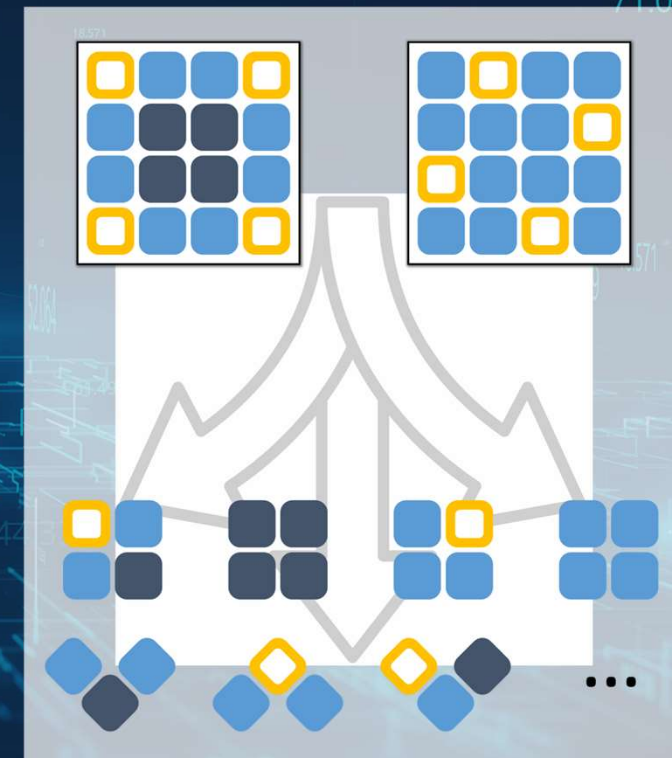
© 2024 Alberto Sangiovanni Vincentelli

Top Down: Extract Common Sub-systems by Partitioning

Partitioning is not enough

- Too many distinct pieces
- $$$$: tape out all of them? 🥺
  - *“We will go bankrupt right away.”, anonymous semiconductor company CEO*





© 2024 Alberto Sangiovanni Vincentelli
Berkeley UNIVERSITY OF CALIFORNIA cadence

Partition and Merge

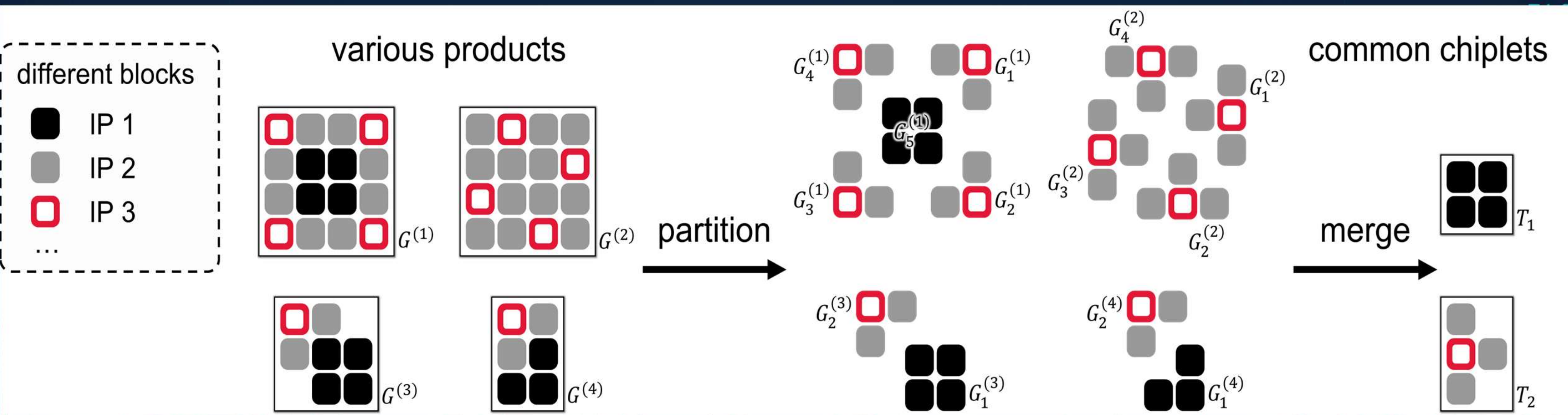
MERGE "similar" pieces later

Construct an overkill to "cover" multiple pieces

- Lower NRE
- Higher BOM
- Total Cost?

© 2024 Alberto Sangiovanni Vincentelli

Chiplet Optimization

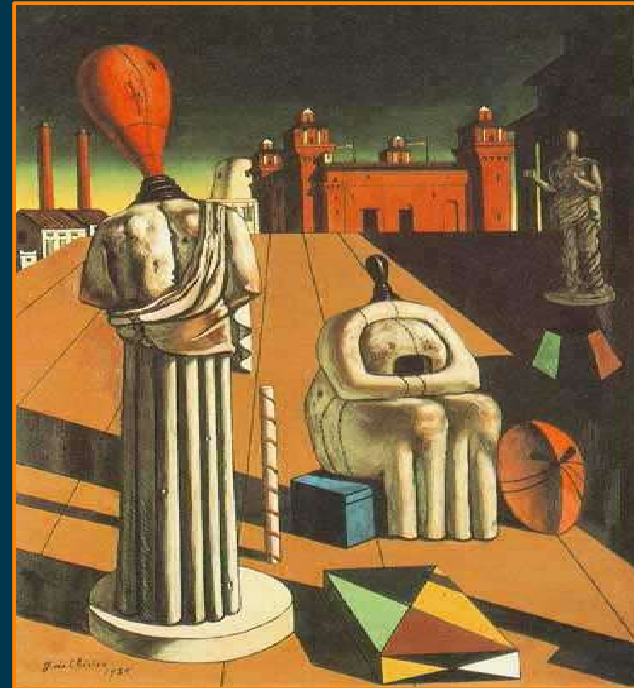


© 2024 Alberto Sangiovanni Vincentelli

Restricted Distribution - Confidential
Berkeley UNIVERSITY OF CALIFORNIA cādence

Final Words of Wisdom

Giuseppe Arcimboldo, The Librarian, 1566
Skokloster Castle, Sweden



Giorgio De Chirico, Le Muse Inquietanti, 1917-18
Collezione Mattioli, Milano, Italy

© 2024 Alberto Sangiovanni Vincentelli

Evolution or Revolution?

- **Natural progression in design methods towards higher level of abstractions and richer set of implementations**
- **Novel approach to implementation to cope with rising costs of integration and difficulty in Moore's law (by the way, Moore predicted chiplets…)**
- **Potential disruption in design and supply chains**
- **Nice problems for EDA researchers!**

© 2024 Alberto Sangiovanni Vincentelli

Outlook

What we need to improve:

- Flow to accelerate the design of chiplets – doing a single SoC is less work than multiple chiplets with the same overall functionality
- Initially, everyone will be convinced they need slightly different chiplets for their system. In the end, they will find the common designs will be sufficient. Similar evolution in the IP space.
- Off-the-shelf available chiplets – be reference designs, GDSII, or even silicon – especially for advanced nodes or community IPs

A compelling cost and TTM story with good enough PPA

© 2024 Alberto Sangiovanni Vincentelli

AI is here to stay… BUT IT IS NOT A PANACEA

- However, foundations are still wobbly
- MANY problems need solutions
- Ethical and geopolitical issues
- Can we inject physics and mathematics into AI Models?

© 2024 Alberto Sangiovanni Vincentelli

Thank You

© 2024 Alberto Sangiovanni Vincentelli