AI in CPS and Semiconductor design: Hype or Reality?

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

1

Outline

- Corsi e ricorsi: Design complexity and Methodologies
- The mega chip trend and limitations
- 3D-IC and Chiplets
- Limitations:
  - Talent
  - Complexity
- Is Machine Learning a Global Solution?
  - Inference implementation: Spiking Neural Networks
  - AlphaFold2
  - Where and Why to apply ML to EDA?
- Concluding Remarks

2

Incipit 40th Design Automation Conference 2003: History repeats itself!

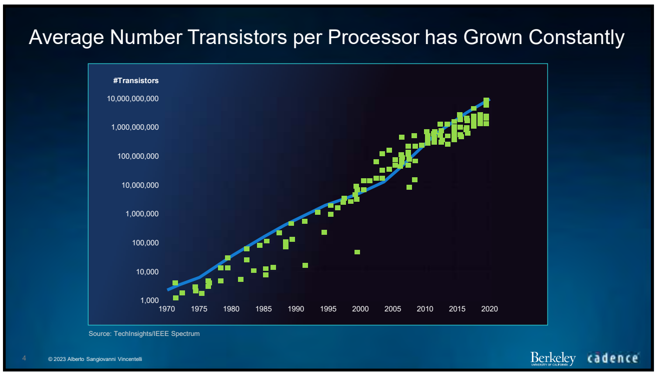
Scientia Nova, De Antiquissima Italorum Sapientia
G.B. Vico (1650)

The Age of the Gods
SENSES

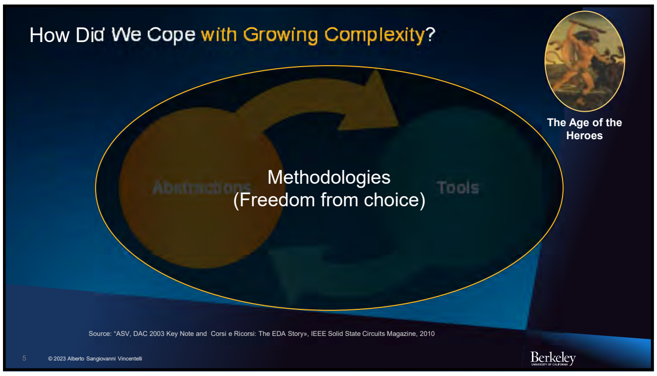
The Age of the Heroes
IMAGINATION

The Age of the Men
REASON

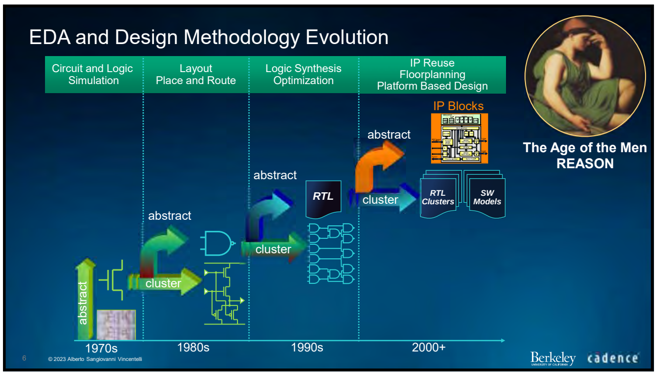
3


Average Number Transistors per Processor has Grown Constantly
#Transistors
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Source: TechInsights/IEEE Spectrum
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

4


How Did We Cope with Growing Complexity?
The Age of the Heroes
Abstractions
Methodologies
(Freedom from choice)
Tools
Source: "ASV, DAC 2003 Key Note and Corsi e Ricorsi: The EDA Story», IEEE Solid State Circuits Magazine, 2010
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

5


EDA and Design Methodology Evolution
Circuit and Logic Simulation
Layout Place and Route
Logic Synthesis Optimization
IP Reuse Floorplanning Platform Based Design
IP Blocks
abstract
cluster
RTL
RTL Clusters
SW Models
The Age of the Men REASON
1970s
1980s
1990s
2000+
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

6

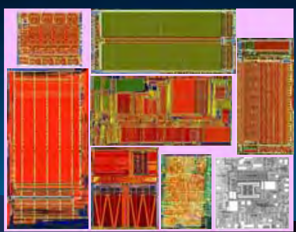
Outline

- Corsi e ricorsi: Design complexity and Methodologies
- **The mega chip trend and limitations**
- 3D-IC and Chiplets
- Limitations:
  - Talent
  - Complexity
- Is Machine Learning a Global Solution?
  - Inference implementation: Spiking Neural Networks
  - AlphaFold2
  - Where and Why to apply ML to EDA?
- Concluding Remarks

Berkeley cadence

7

IP Based Design:
Enabling System Companies to Design Chips
(Amazon, Apple, Google, Microsoft, Tesla….)



- **Assemble Components from parameterized library**

  Including:
  - **Configurable processor core**
  - **Memories (RAM, ROM)**
  - **Special-purpose standard blocks (ASSPs)**
  - **Glue Logic**
  - **Third-party special-purpose logic/MEMS/MEOS**
- **Integrate using standard approach to on-chip communication**

Berkeley

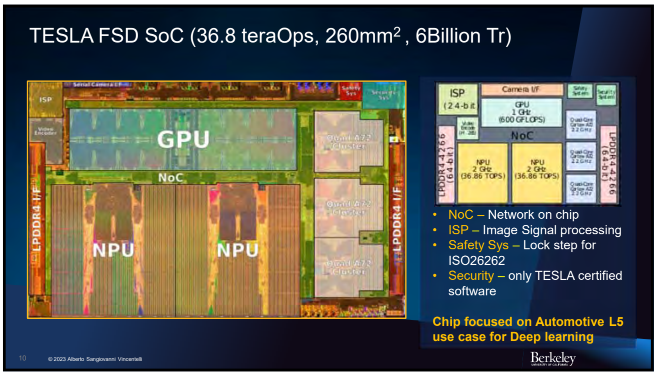
8

Apple Monster Chip

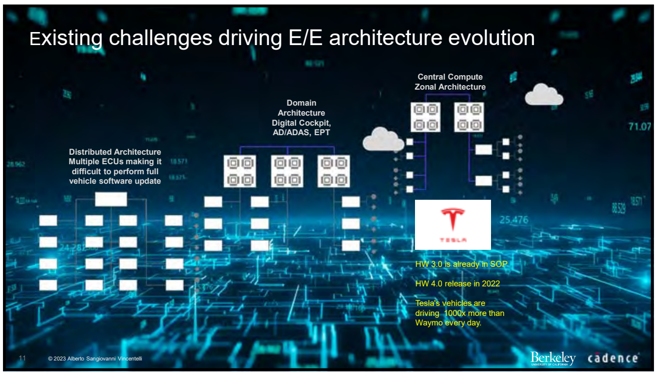
Apple A11 Bionic
4.3Billion transistors
34 GOPS
87.66 mm2

Berkeley

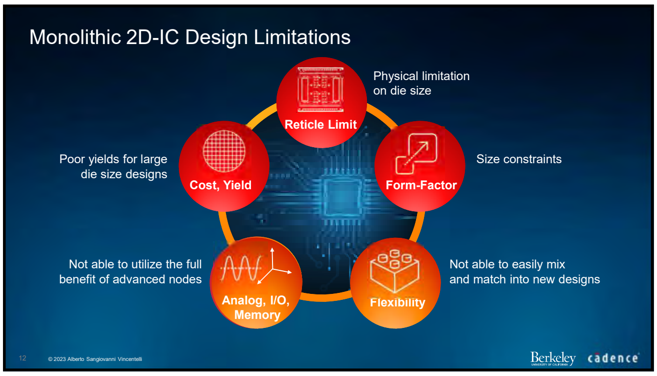
9


TESLA FSD SoC (36.8 teraOps, 260mm2 , 6Billion Tr)
GPU
NoC
NPU
NPU
LPDDR4 I/F
LPDDR4 I/F
ISP
Camera I/F
NoC
NPU 2 GHz (36.86 TOPS)
NPU 2 GHz (36.86 TOPS)
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
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

10


Existing challenges driving E/E architecture evolution
Central Compute Zonal Architecture
Domain Architecture Digital Cockpit, AD/ADAS, EPT
Distributed Architecture Multiple ECUs making it difficult to perform full vehicle software update
HW 3.0 is already in SOP
HW 4.0 release in 2022
Tesla's vehicles are driving 1000x more than Waymo every day
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

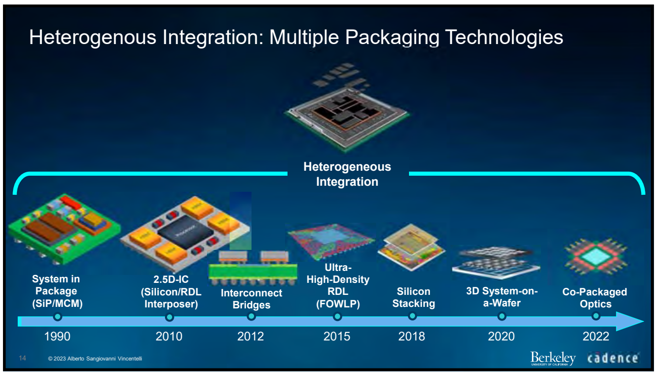
11


Monolithic 2D-IC Design Limitations
Physical limitation on die size
Reticle Limit
Poor yields for large die size designs
Cost, Yield
Size constraints
Form-Factor
Not able to utilize the full benefit of advanced nodes
Analog, I/O, Memory
Not able to easily mix and match into new designs
Flexibility
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

12


Outline
Corsi e ricorsi: Design complexity and Methodologies
The mega chip trend and limitations
3D-IC and Chiplets
Limitations:
Talent
Complexity
Is Machine Learning a Global Solution?
Inference implementation: Spiking Neural Netwroks
AlphaFold2
Where and Why to apply ML to EDA?
Concluding Remarks
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

13

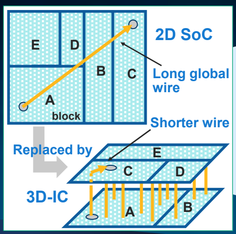

Heterogenous Integration: Multiple Packaging Technologies
Heterogeneous Integration
System in Package (SiP/MCM)
2.5D-IC (Silicon/RDL Interposer)
Interconnect Bridges
Ultra-High-Density RDL (FOWLP)
Silicon Stacking
3D System-on-a-Wafer
Co-Packaged Optics
1990
2010
2012
2015
2018
2020
2022
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

14


Corsi e Ricorsi:
Déjà vu all over again?
MCM Start-ups in the 80s:
Polycon, Advanced Packaging Systems, ISA, Polylithics, Alcoa Microelectronics, nChip,
TIOs HDI and Pacific Microelectronics Center
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

15

Not really....Silicon Stacking (3D-IC)

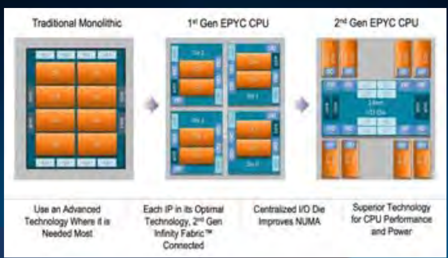


- Shorter Wire
- Less Power
- Higher Performance
- Higher Bandwidth
- Smaller Profile
- Better Yield

Berkeley cadence

16

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

Berkeley

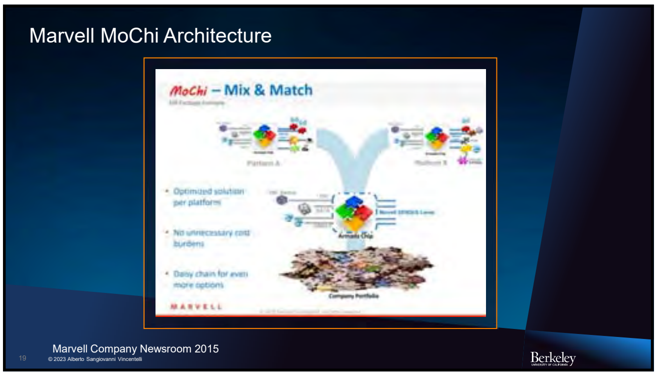
17

Intel SHIP Program (April 2023)

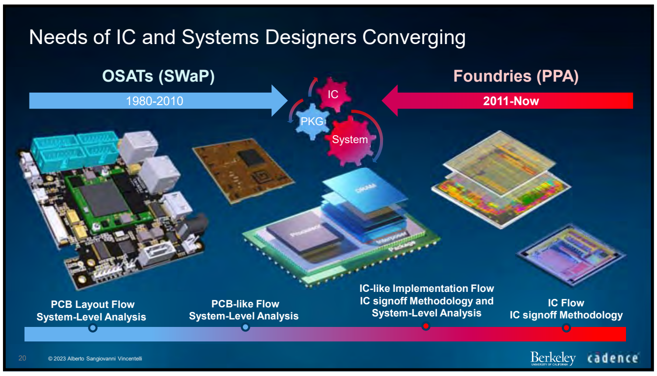
Intel Corporation Newsroom

Berkeley

18


Marvell MoChi Architecture
MoChi – Mix & Match
Optimized solution per platform
No unnecessary cost burdens
Daisy chain for even more options
MARVELL
Marvell Company Newsroom 2015
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

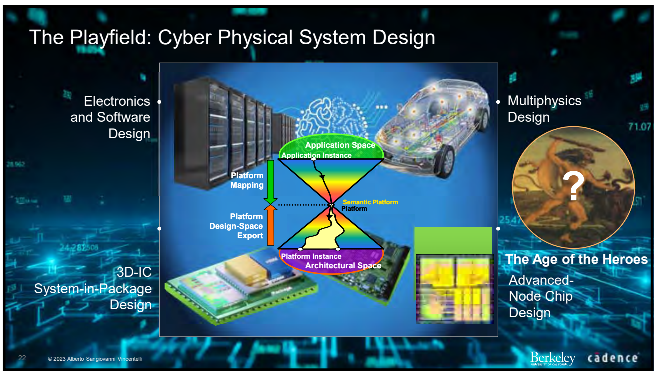
19


Needs of IC and Systems Designers Converging
OSATs (SWaP)
1980-2010
Foundries (PPA)
2011-Now
IC
PKG
System
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
Berkeley
cadence

20


Cyber Physical Systems
Interconnect the World Around Us and Make It "Smarter"
Instrumented
Interconnected
Intelligent
The Age of the Gods
Autonomous Driving
Life Sciences
Telecommunications
Avionics
Transportation
Military systems
Power generation
and distribution
Factory
automation
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

21


The Playfield: Cyber Physical System Design
Electronics and Software Design
Multiphysics Design
Application Space
Application Instance
Platform Mapping
Semantic Platform
Platform Design-Space Export
Platform Instance
Architectural Space
?
The Age of the Heroes
3D-IC System-in-Package Design
Advanced-Node Chip Design
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

22

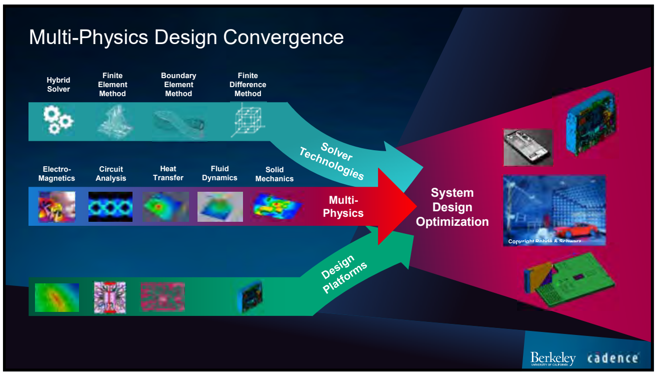

Outline
Corsi e ricorsi: Design complexity and Methodologies
The mega chip trend and limitations
3D-IC and Chiplets
Limitations:
Complexity
Is Machine Learning a Global Solution?
Inference implementation: Spiking Neural Networks
AlphaFold2
Where and Why to apply ML to EDA?
Concluding Remarks
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

23


Outline
Corsi e ricorsi: Design complexity and Methodologies
The mega chip trend and limitations
3D-IC and Chiplets
Limitations:
Complexity
Talent
Is Machine Learning a Global Solution?
Inference implementation: Spiking Neural Networks
AlphaFold2
Where and Why to apply ML to EDA?
Concluding Remarks
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

24

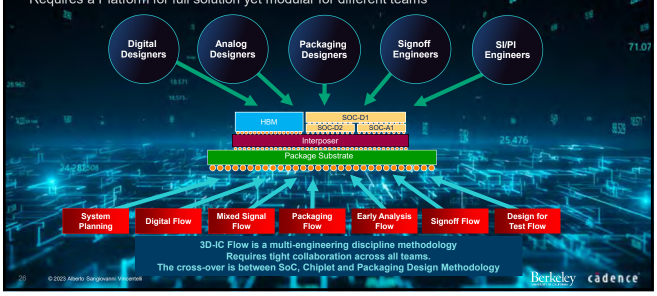
Multi-Physics Design Convergence



25

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

Requires a Platform for full solution yet modular for different teams



26

Incredible Complexity: Number and type of components

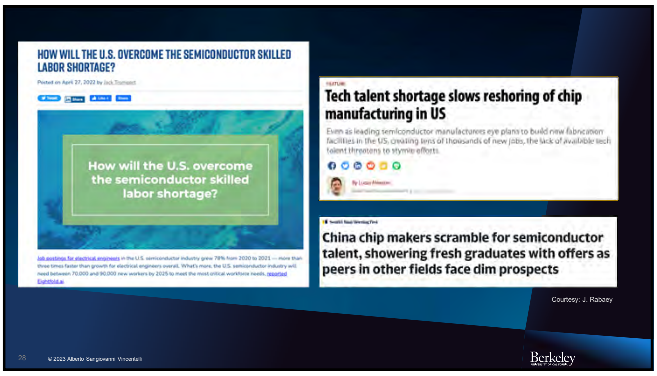
Huge design space: Exploration problem

Verification

Lack of talent

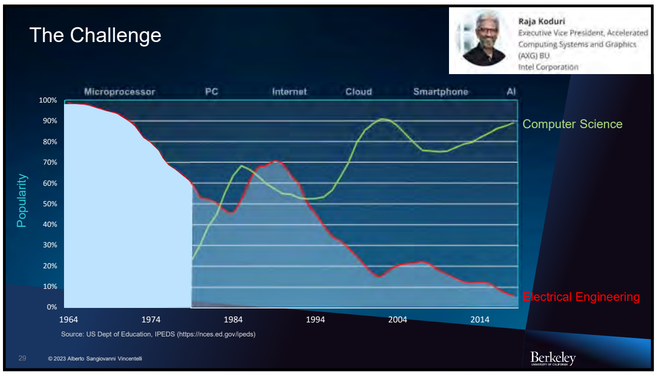
Berkeley cadence

27



HOW WILL THE U.S. OVERCOME THE SEMICONDUCTOR SKILLED LABOR SHORTAGE?
Posted on April 27, 2022 by Jack Trompert
How will the U.S. overcome the semiconductor skilled labor shortage?
Job postings for electrical engineers in the U.S. semiconductor industry grew 78% from 2020 to 2021 — more than three times faster than growth for electrical engineers overall. What's more, the U.S. semiconductor industry will need between 70,000 and 90,000 new workers by 2025 to meet the most critical workforce needs, reported Eightfold.ai
Tech talent shortage slows reshoring of chip manufacturing in US
Even as leading semiconductor manufacturers eye plans to build new fabrication facilities in the US, creating tens of thousands of new jobs, the lack of available tech talent threatens to stymie efforts.
China chip makers scramble for semiconductor talent, showering fresh graduates with offers as peers in other fields face dim prospects
Courtesy: J. Rabaey
28
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

28



The Challenge
Raja Koduri
Executive Vice President, Accelerated Computing Systems and Graphics (AXG) BU
Intel Corporation
Microprocessor
PC
Internet
Cloud
Smartphone
AI
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Popularity
1964
1974
1984
1994
2004
2014
Computer Science
Electrical Engineering
Source: US Dept of Education, IPEDS (https://nces.ed.gov/ipeds)
29
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

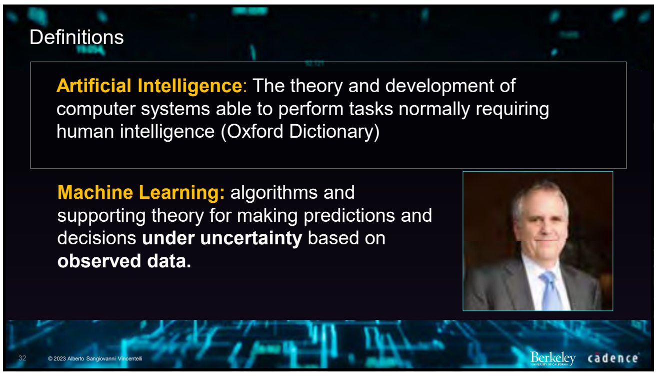
29

Outline
Corsi e ricorsi: Design complexity and Methodologies
The mega chip trend and limitations
3D-IC and Chiplets
Limitations:
Talent
Complexity
Is Machine Learning a Global Solution?
Inference implementation: Spiking Neural Networks
AlphaFold2
Where and Why to apply ML to EDA?
Concluding Remarks
30
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

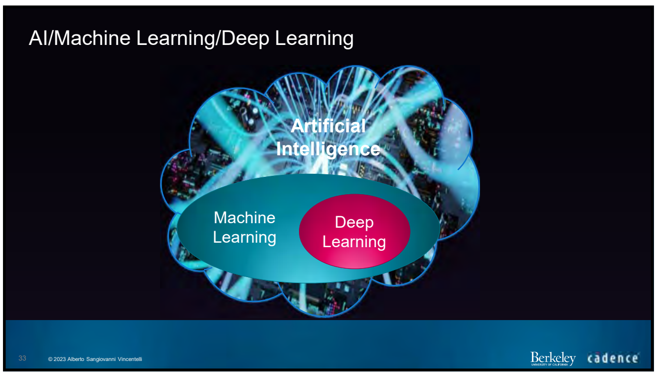
30


Is AI a Panacea?
171
Talks
1
Workshop
6
Tutorials
AI
6
Keynotes and visionary talks
2
Panels
24
Sessions
@2023 DAC
©2023 Alberto Sangiovanni Vincentelli

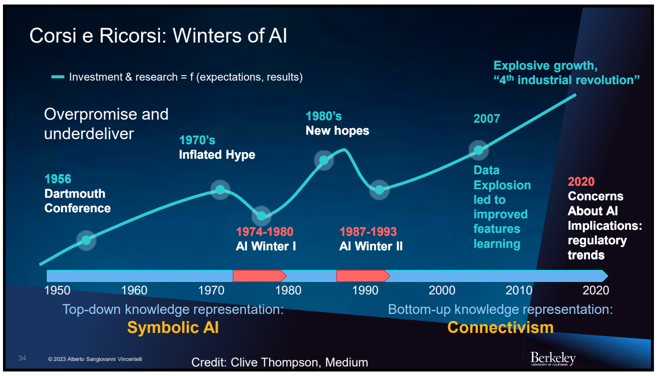
31


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.
©2023 Alberto Sangiovanni Vincentelli

32

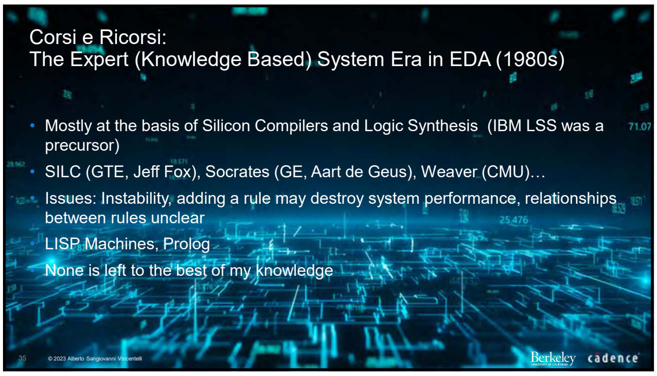

AI/Machine Learning/Deep Learning
Artificial Intelligence
Machine Learning
Deep Learning
©2023 Alberto Sangiovanni Vincentelli

33



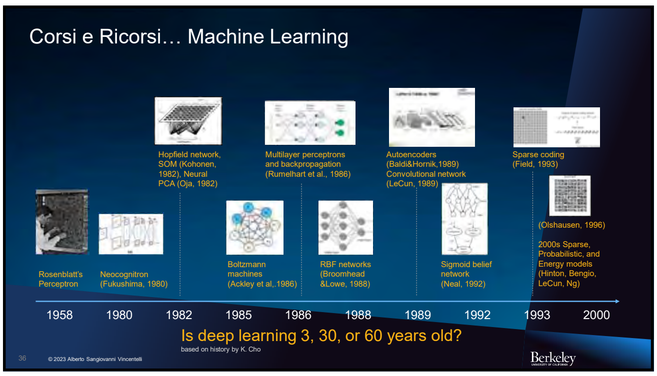
Corsi e Ricorsi: Winters of AI
Investment & research = f (expectations, results)
Overpromise and underdeliver
1956 Dartmouth Conference
1970's Inflated Hype
1974-1980 AI Winter I
1980's New hopes
1987-1993 AI Winter II
2007
Data Explosion led to improved features learning
Explosive growth, "4th industrial revolution"
2020 Concerns About AI Implications: regulatory trends
1950
1960
1970
1980
1990
2000
2010
2020
Top-down knowledge representation:
Symbolic AI
Bottom-up knowledge representation:
Connectivism
34
© 2023 Alberto Sangiovanni Vincentelli
Credit: Clive Thompson, Medium
Berkeley

34



Corsi e Ricorsi:
The Expert (Knowledge Based) System Era in EDA (1980s)
Mostly at the basis of Silicon Compilers and Logic Synthesis (IBM LSS was a precursor)
SILC (GTE, Jeff Fox), Socrates (GE, Aart de Geus), Weaver (CMU)…
Issues: Instability, adding a rule may destroy system performance, relationships between rules unclear
LISP Machines, Prolog
None is left to the best of my knowledge
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

35



Corsi e Ricorsi… 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)
(Olshausen, 1996)
2000s Sparse, Probabilistic, and Energy models (Hinton, Bengio, LeCun, Ng)
Rosenblatt's Perceptron
Neocognitron (Fukushima, 1980)
Boltzmann machines (Ackley et al,.1986)
RBF networks (Broomhead &Lowe, 1988)
Sigmoid belief network (Neal, 1992)
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
36
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

36

ASV and AI

- Alan Kramer and Alberto Sangiovanni-Vincentelli, Efficient Parallel Learning Algorithms for Neural Networks, Proceedings of the IEEE Conference on Neural Information Processing, Denver CO, 1989.
- Alan Kramer, P. Ko, and Alberto Sangiovanni-Vincentelli, Massively Parallel Analog Geometric Computation Using EEPROMS, Neural Networks for Computing Conference (abstracts), Snowbird UT, Apr. 1991.
- C. K. Sin, Alan Kramer, V. Hu, R. Chu, P. Ko, and Alberto Sangiovanni-Vincentelli, EEPROM as an Analog Storage Device with Particular Application in Neural Networks, IEEE Transactions on Electron Devices, Vol. 39, No. 6, pp. 1410-1419, Jun. 1992.

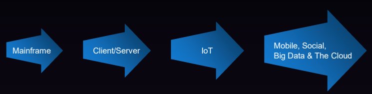
37

Learning and Boolean Functions

- Arlindo Oliveira and Alberto Sangiovanni-Vincentelli, Learning Concepts by Synthesizing Minimal Threshold Gate Networks, Proceedings of the Eigth International Workshop in Machine Learning, Chicago IL, pp. 193-197, 1991.
- Arlindo Oliveira and Alberto Sangiovanni-Vincentelli, LSAT - An Algorithm for the Synthesis of Two Level Threshold Gate Networks, Proceedings IEEE International Conference on Computer Aided Design (ICCAD-91), Santa Clara CA, pp. 130-133, Nov. 1991.
- Arlindo Oliveira and Alberto Sangiovanni-Vincentelli, Synthesis of Minimal Multi-Level Networks, Neural Networks for Computing Conference (abstracts), Snowbird UT, Apr. 1992.
- Arlindo Oliveira and Alberto Sangiovanni-Vincentelli, Constructive Induction Using a Non-Greedy Strategy for Feature Selection, Proceedings of the Ninth International Conference in Machine Learning, Scotland UK, pp. 355-360, Jul. 1992.
- Arlindo Oliveira and Alberto Sangiovanni-Vincentelli, What Can Boolean Networks Learn?, Proceedings of the 3rd International Workshop on Computational Learning Theory and Natural Learning Systems, Madison WI, Aug. 1992.
- Arlindo Oliveira and Alberto Sangiovanni-Vincentelli, Learning Complex Boolean Functions: Algorithms and Applications, Proceedings of Neural Information Processing Systems Conference, Denver CO, Dec. 1993.

38

BIG data



Every 60 seconds+

98,000+ tweets

695,000 status updates

11 million instant messages

698,445 Google searches

168 million_+ emails sent

1,820TB of data created

217 new mobile web uses

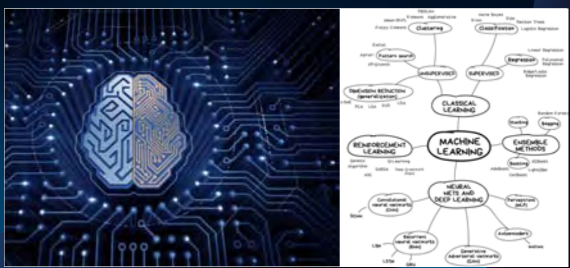
Source: Big Data: the Future Has Arrived 2016

39

Big Data + Processing Power =
New Age for Artificial Intelligence

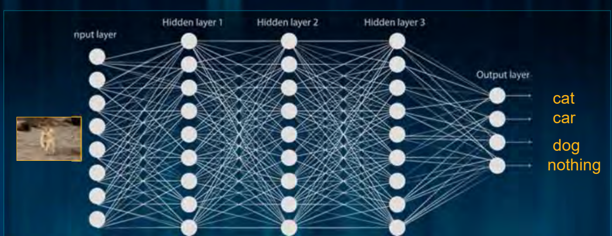
40

Machine Learning Galaxy



41

Deep Learning: Many Layer Neural Network



different weights = different computation

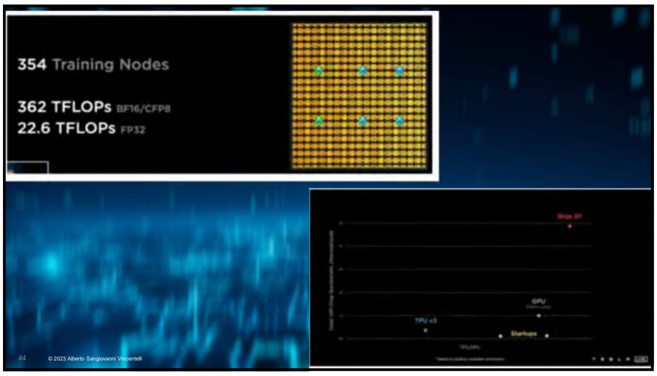
Neural Net Training: Find the weights that minimize the difference between labels and activation.

42

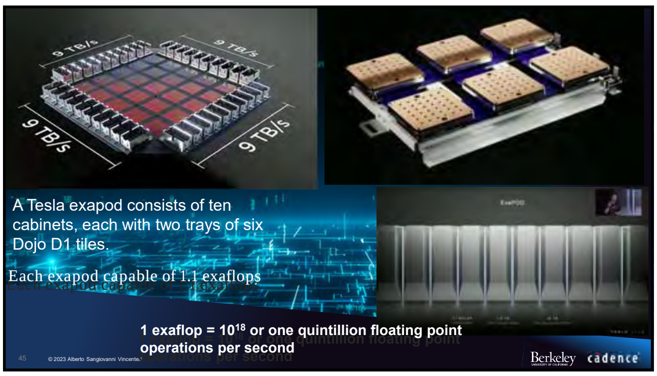
Machine Learning and Approximations

- 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?

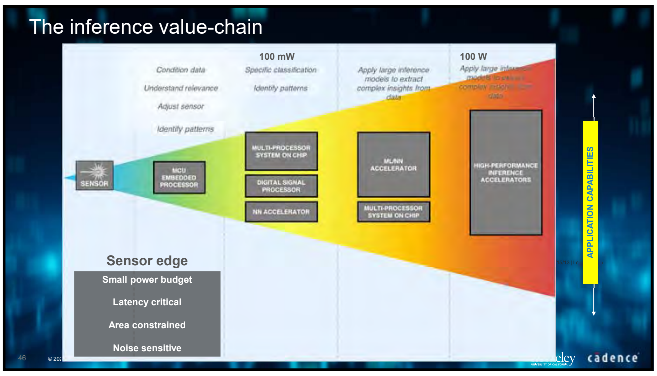
43



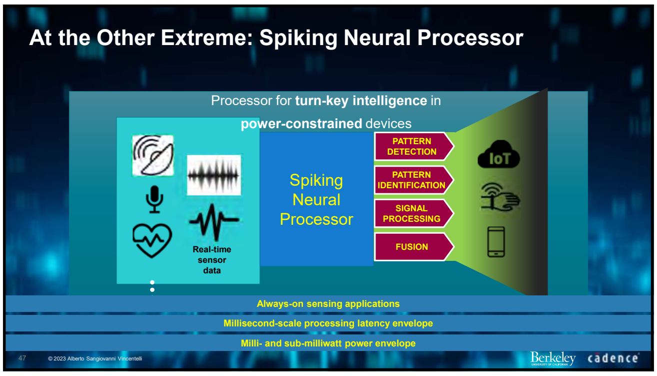
44



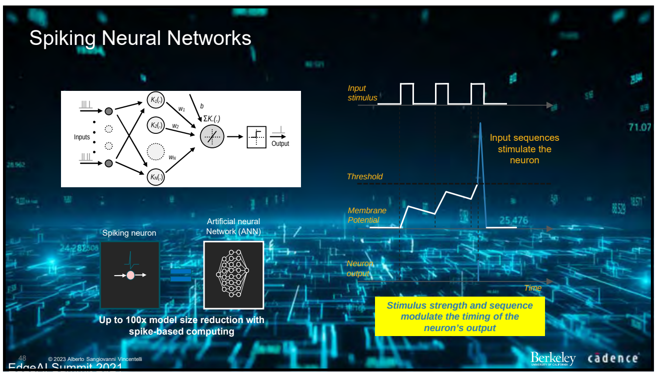
45


The inference value-chain
100 mW
100 W
Condition data
Understand relevance
Adjust sensor
Identify patterns
Specific classification
Identify patterns
Apply large inference models to extract complex insights from data
SENSOR
MCU EMBEDDED PROCESSOR
MULTI-PROCESSOR SYSTEM ON CHIP
DIGITAL SIGNAL PROCESSOR
NN ACCELERATOR
ML/NN ACCELERATOR
MULTI-PROCESSOR SYSTEM ON CHIP
HIGH-PERFORMANCE INFERENCE ACCELERATORS
APPLICATION CAPABILITIES
Sensor edge
Small power budget
Latency critical
Area constrained
Noise sensitive
46
cadence

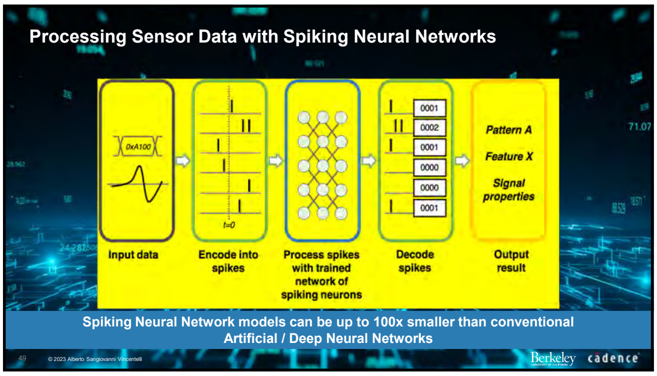
46


At the Other Extreme: Spiking Neural Processor
Processor for turn-key intelligence in power-constrained devices
Spiking Neural Processor
PATTERN DETECTION
PATTERN IDENTIFICATION
SIGNAL PROCESSING
FUSION
IoT
Real-time sensor data
Always-on sensing applications
Millisecond-scale processing latency envelope
Milli- and sub-milliwatt power envelope
47
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

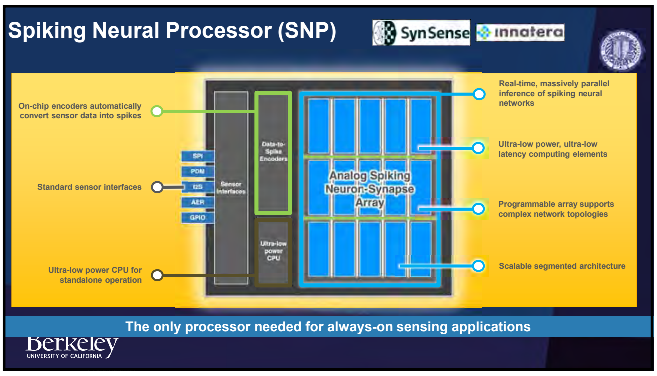
47


Spiking Neural Networks
Inputs
Output
Input stimulus
Input sequences stimulate the neuron
Threshold
Membrane Potential
Neuron output
Time
Spiking neuron
Artificial neural Network (ANN)
Up to 100x model size reduction with spike-based computing
Stimulus strength and sequence modulate the timing of the neuron's output
48
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

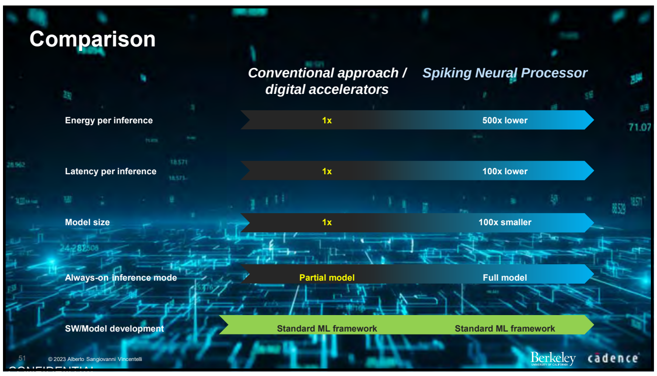
48


Processing Sensor Data with Spiking Neural Networks
0xA100
t=0
0001
0002
0001
0000
0000
0001
Pattern A
Feature X
Signal properties
Input data
Encode into spikes
Process spikes with trained network of spiking neurons
Decode spikes
Output result
Spiking Neural Network models can be up to 100x smaller than conventional Artificial / Deep Neural Networks
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

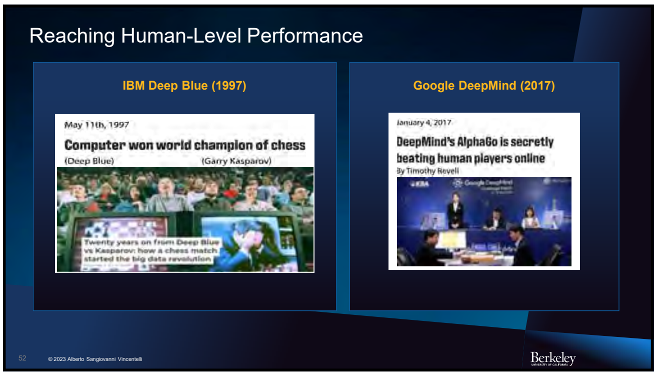
49


Spiking Neural Processor (SNP)
SynSense
innatera
On-chip encoders automatically convert sensor data into spikes
Standard sensor interfaces
Ultra-low power CPU for standalone operation
SPI
PDM
I2S
AER
GPIO
Sensor Interfaces
Data-to-Spike Encoders
Ultra-low power CPU
Analog Spiking Neuron-Synapse Array
Real-time, massively parallel inference of spiking neural networks
Ultra-low power, ultra-low latency computing elements
Programmable array supports complex network topologies
Scalable segmented architecture
The only processor needed for always-on sensing applications
Berkeley
UNIVERSITY OF CALIFORNIA

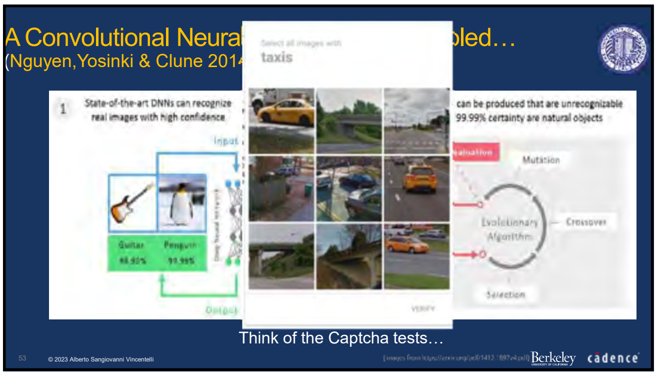
50


Comparison
Conventional approach / digital accelerators
Spiking Neural Processor
Energy per inference
1x
500x lower
Latency per inference
1x
100x lower
Model size
1x
100x smaller
Always-on inference mode
Partial model
Full model
SW/Model development
Standard ML framework
Standard ML framework
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

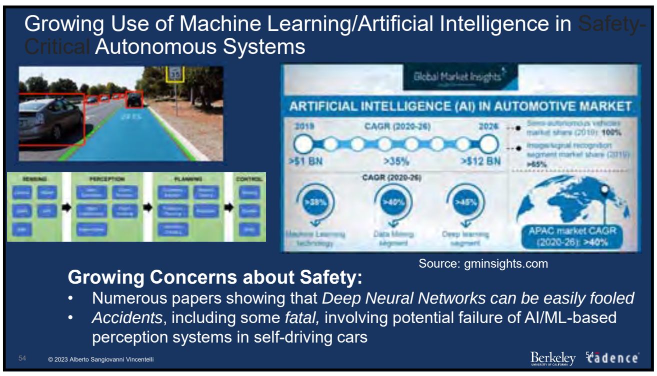
51


Reaching Human-Level Performance
IBM Deep Blue (1997)
May 11th, 1997
Computer won world champion of chess
(Deep Blue)
(Garry Kasparov)
Twenty years on from Deep Blue vs Kasparov: how a chess match started the big data revolution
Google DeepMind (2017)
January 4, 2017
DeepMind's AlphaGo is secretly beating human players online
By Timothy Revell
52
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

52


A Convolutional Neural ... fooled…
(Nguyen,Yosinki & Clune 2014
Select all images with
taxis
State-of-the-art DNNs can recognize real images with high confidence
Input
Guitar
Penguin
Output
can be produced that are unrecognizable 99.99% certainty are natural objects
Mutation
Evolutionary Algorithm
Crossover
Selection
VERIFY
Think of the Captcha tests…
53
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

53


Growing Use of Machine Learning/Artificial Intelligence in Safety-Critical Autonomous Systems
Global Market Insights
ARTIFICIAL INTELLIGENCE (AI) IN AUTOMOTIVE MARKET
2019
CAGR (2020-26)
2026
>$1 BN
>35%
>$12 BN
CAGR (2020-26)
APAC market CAGR (2020-26): >40%
Source: gminsights.com
Growing Concerns about Safety:
• Numerous papers showing that *Deep Neural Networks can be easily fooled*
• *Accidents*, including some *fatal,* involving potential failure of AI/ML-based perception systems in self-driving cars
54
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

54

IMPORTANT Result: Determining 3-D Protein Structure AlphaFold and AlphaFold 2

The problem: given a protein's amino acid sequence, what is its three-dimensional atomic structure?

The 3-D structure of an enzyme from the bacteria Colwellia psychrerythraea

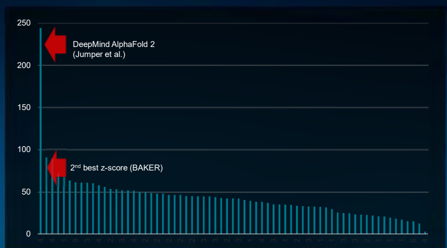
Berkeley cadence

55

Google DeepMind's AlphaFold 2 Neural Network Breakthrough

Computational biologists' predict structure of a protein from its sequence

Critical Assessment of Structural Prediction (CASP) competition, a biannual blind test where computational biologists try to predict the structure of several proteins whose structure has been determined experimentally — yet not publicly released.

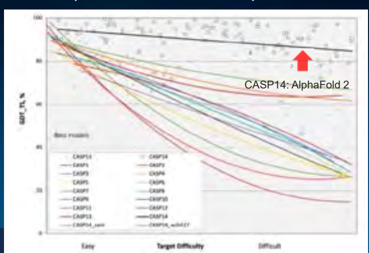


Source: Oxford Protein Informatics Group (OPIG), Critical Assessment of Techniques for Protein Structure Prediction (CASP), December 1 , 2020.
Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Berkeley cadence

56

History of Critical Assessment of Techniques for Protein Structure Prediction (CASP1 to CASP14)



GDT_TS around 60% represents a "correct fold", we have an idea of how the protein folds globally; and over 80% we start seeing side chains that closely resemble the model.

Courtesy: Oxford Protein Informatics Group (OPIG), Critical Assessment of Techniques for Protein Structure Prediction (CASP), December 1 , 2020.
Graph by John Moult, chair of Oxford Informatics Group Article.

Berkeley

57

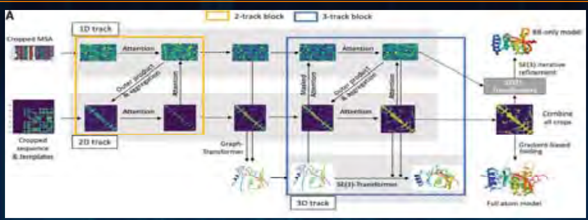
How did they do it?

ASV Jumper Baker

Berkeley

58

Deep Learning Architecture: leveraging **deep** knowledge!



Baker et al., SCIENCE, 2021 Vol 373, Issue 6557 pp. 871-876

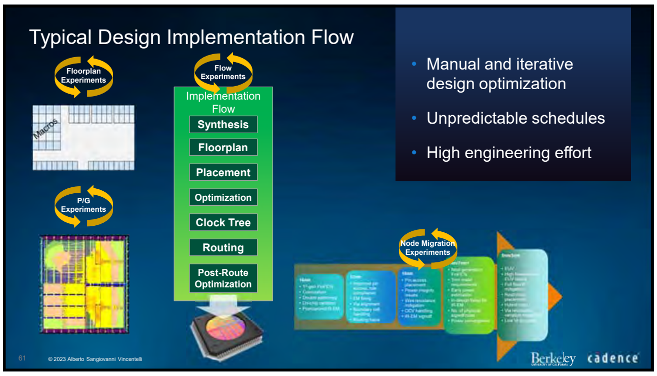
Berkeley cadence

59

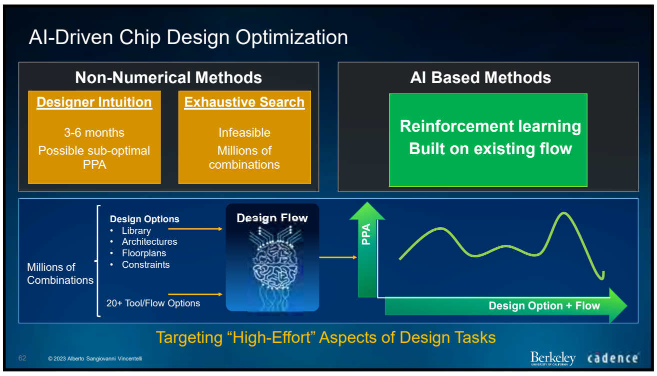
Where and How to Use ML in Design?

Berkeley cadence

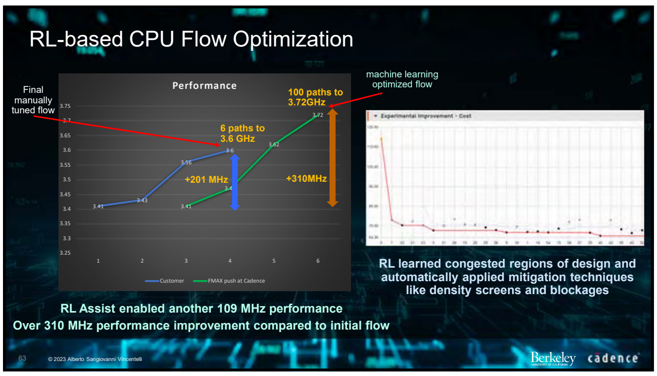
60


Typical Design Implementation Flow
Floorplan Experiments
Flow Experiments
Implementation Flow
Synthesis
Floorplan
Placement
Optimization
Clock Tree
Routing
Post-Route Optimization
P/G Experiments
Node Migration Experiments
Manual and iterative design optimization
Unpredictable schedules
High engineering effort
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

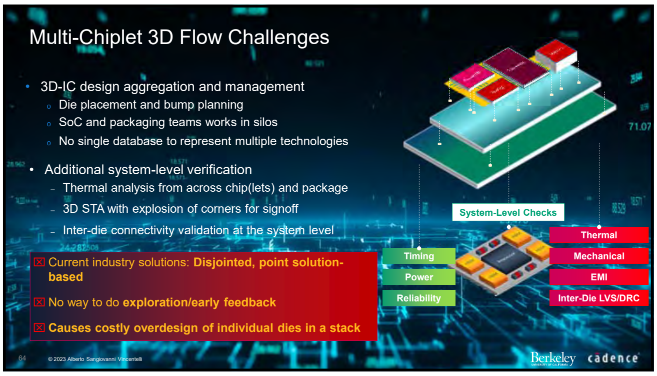
61


AI-Driven Chip Design Optimization
Non-Numerical Methods
Designer Intuition
3-6 months
Possible sub-optimal PPA
Exhaustive Search
Infeasible
Millions of combinations
AI Based Methods
Reinforcement learning
Built on existing flow
Millions of Combinations
Design Options
Library
Architectures
Floorplans
Constraints
20+ Tool/Flow Options
Design Flow
PPA
Design Option + Flow
Targeting "High-Effort" Aspects of Design Tasks
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

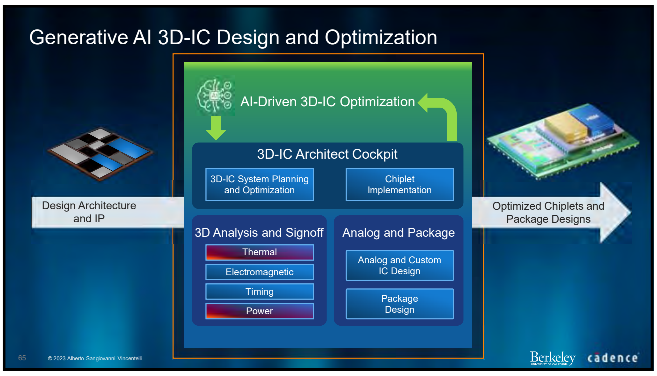
62


RL-based CPU Flow Optimization
Final manually tuned flow
Performance
100 paths to 3.72GHz
6 paths to 3.6 GHz
+201 MHz
+310MHz
machine learning optimized flow
RL learned congested regions of design and automatically applied mitigation techniques like density screens and blockages
RL Assist enabled another 109 MHz performance
Over 310 MHz performance improvement compared to initial flow
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

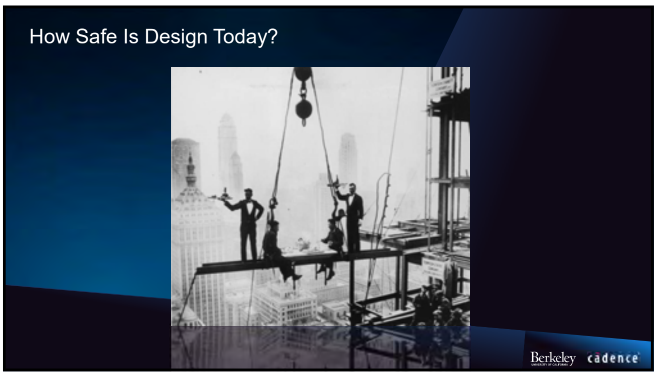
63


Multi-Chiplet 3D Flow Challenges
3D-IC design aggregation and management
Die placement and bump planning
SoC and packaging teams works in silos
No single database to represent multiple technologies
Additional system-level verification
Thermal analysis from across chip(lets) and package
3D STA with explosion of corners for signoff
Inter-die connectivity validation at the system level
Current industry solutions: Disjointed, point solution-based
No way to do exploration/early feedback
Causes costly overdesign of individual dies in a stack
System-Level Checks
Timing
Power
Reliability
Thermal
Mechanical
EMI
Inter-Die LVS/DRC
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

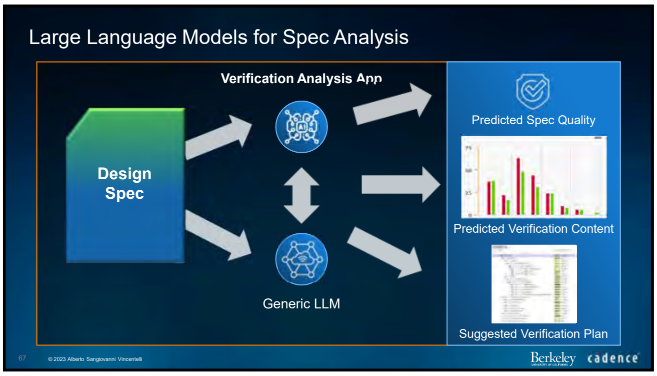
64


Generative AI 3D-IC Design and Optimization
AI-Driven 3D-IC Optimization
3D-IC Architect Cockpit
3D-IC System Planning and Optimization
Chiplet Implementation
Design Architecture and IP
Optimized Chiplets and Package Designs
3D Analysis and Signoff
Thermal
Electromagnetic
Timing
Power
Analog and Package
Analog and Custom IC Design
Package Design
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

65

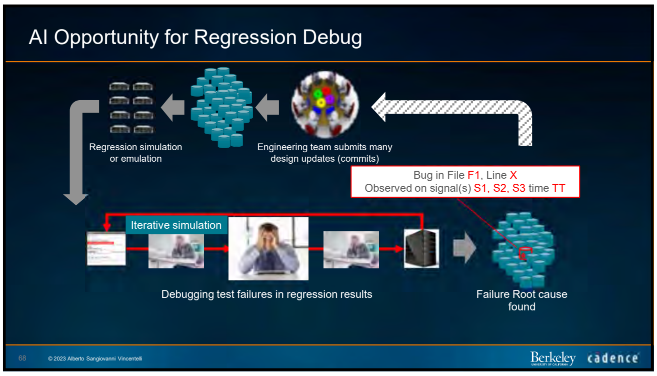

How Safe Is Design Today?
Berkeley
cadence

66



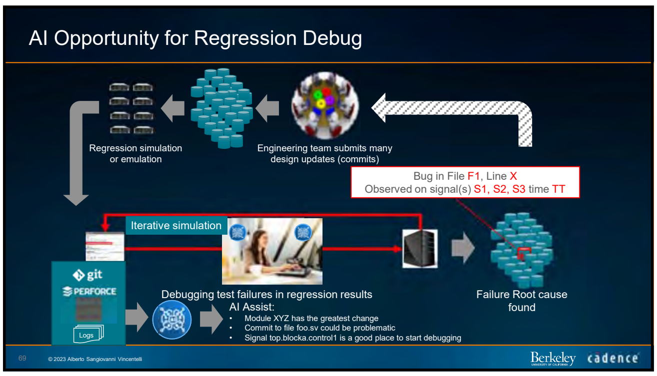
Large Language Models for Spec Analysis
Verification Analysis App
Design Spec
Generic LLM
Predicted Spec Quality
Predicted Verification Content
Suggested Verification Plan
67
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

67



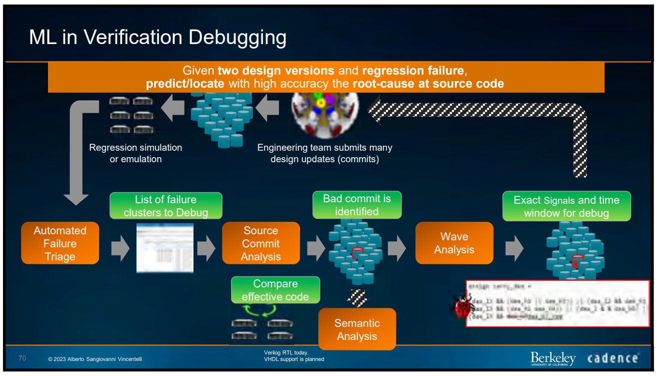
AI Opportunity for Regression Debug
Regression simulation or emulation
Engineering team submits many design updates (commits)
Bug in File F1, Line X
Observed on signal(s) S1, S2, S3 time TT
Iterative simulation
Debugging test failures in regression results
Failure Root cause found
68
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

68



AI Opportunity for Regression Debug
Regression simulation or emulation
Engineering team submits many design updates (commits)
Bug in File F1, Line X
Observed on signal(s) S1, S2, S3 time TT
Iterative simulation
git
PERFORCE
Logs
Debugging test failures in regression results
AI Assist:
• Module XYZ has the greatest change
• Commit to file foo.sv could be problematic
• Signal top.blocka.control1 is a good place to start debugging
Failure Root cause found
69
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

69

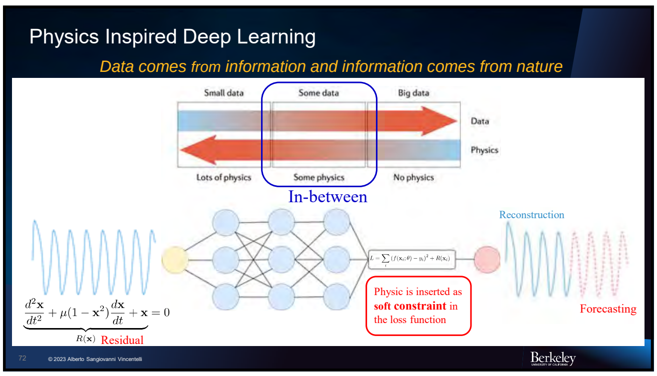


ML in Verification Debugging
Given two design versions and regression failure, predict/locate with high accuracy the root-cause at source code
Regression simulation or emulation
Engineering team submits many design updates (commits)
Automated Failure Triage
List of failure clusters to Debug
Source Commit Analysis
Bad commit is identified
Wave Analysis
Exact Signals and time window for debug
Compare effective code
Semantic Analysis
70
© 2023 Alberto Sangiovanni Vincentelli
Verilog RTL today
VHDL support is planned
Berkeley
cadence

70

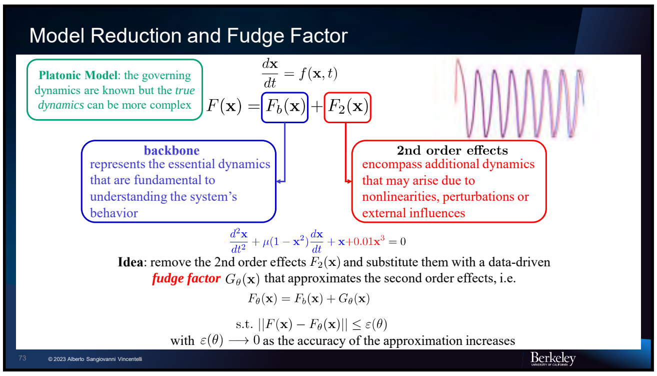
Final Words of Wisdom
Giuseppe Arcimboldo, The Librarian, 1566
Skokloster Castle, Sweden
Giorgio De Chirico, Le Muse Inquietanti, 1917-18
Collezione Mattioli, Milano, Italy
71
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

71

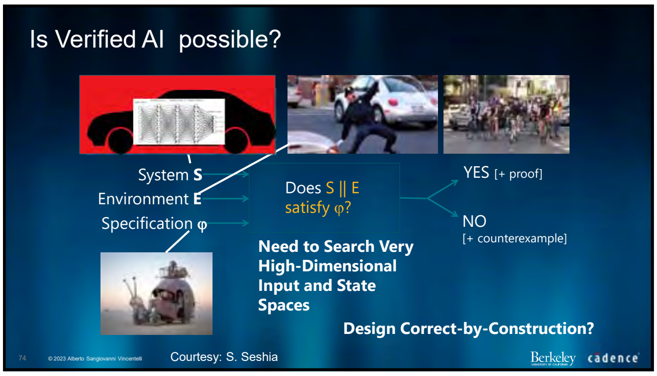


Physics Inspired Deep Learning
Data comes from information and information comes from nature
Small data
Some data
Big data
Data
Physics
Lots of physics
Some physics
No physics
In-between
Reconstruction
Physic is inserted as soft constraint in the loss function
Forecasting
$\frac{d^2x}{dt^2} + \mu(1 - x^2)\frac{dx}{dt} + x = 0$
$R(x)$ Residual
72
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

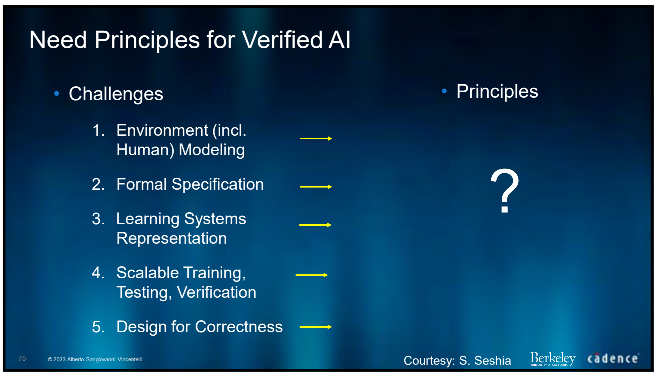
72


Model Reduction and Fudge Factor
Platonic Model: the governing dynamics are known but the true dynamics can be more complex
$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, t)$
$F(\mathbf{x}) = \boxed{F_b(\mathbf{x})} + \boxed{F_2(\mathbf{x})}$
backbone
represents the essential dynamics that are fundamental to understanding the system's behavior
2nd order effects
encompass additional dynamics that may arise due to nonlinearities, perturbations or external influences
$\frac{d^2\mathbf{x}}{dt^2} + \mu(1-\mathbf{x}^2)\frac{d\mathbf{x}}{dt} + \mathbf{x} + 0.01\mathbf{x}^3 = 0$
Idea: remove the 2nd order effects $F_2(\mathbf{x})$ and substitute them with a data-driven fudge factor $G_\theta(\mathbf{x})$ that approximates the second order effects, i.e.
$F_\theta(\mathbf{x}) = F_b(\mathbf{x}) + G_\theta(\mathbf{x})$
s.t. $||F(\mathbf{x}) - F_\theta(\mathbf{x})|| \leq \varepsilon(\theta)$
with $\varepsilon(\theta) \longrightarrow 0$ as the accuracy of the approximation increases
73
© 2023 Alberto Sangiovanni Vincentelli
Berkeley

73


Is Verified AI possible?
System S
Environment E
Specification φ
Does S || E satisfy φ?
YES [+ proof]
NO [+ counterexample]
Need to Search Very High-Dimensional Input and State Spaces
Design Correct-by-Construction?
74
© 2023 Alberto Sangiovanni Vincentelli
Courtesy: S. Seshia
Berkeley
cadence

74


Need Principles for Verified AI
Challenges
1. Environment (incl. Human) Modeling
2. Formal Specification
3. Learning Systems Representation
4. Scalable Training, Testing, Verification
5. Design for Correctness
Principles
?
75
© 2023 Alberto Sangiovanni Vincentelli
Courtesy: S. Seshia
Berkeley
cadence

75

Concerns Around AI Biases are Mounting

AI transparency tech, also known as explainable AI, traces back outputs from AI algorithms to provide a way to understand what's happening in "human terms."

As AI is increasingly used for decision-making across industries, understanding how and why an algorithm makes its decisions can help mitigate inherent biases associated with most AI systems in existence today.

McKinsey Report on AI

Berkeley

76

Thank You

Berkeley cadence

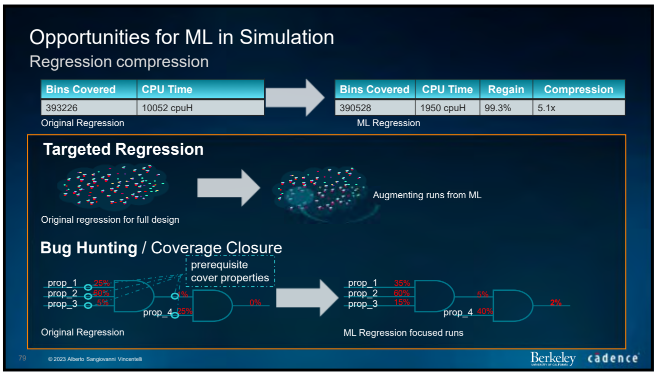
77

Long time ago in a country far away….

- Alberto Sangiovanni-Vincentelli and Mauro Somalvico, State Space Approach in Problem Solving Optimization, Optimization Techniques, P. Conti and E. Ruberti, Editors, Springer-Verlag , New York NY, 1973.
- Alberto Sangiovanni-Vincentelli and Mauro Somalvico, Problem Solving Methods in Computer Aided Medical Diagnosis, in Proceedings of the 20th International Electrical Congress on Electronics, Rome, Italy, pp. 28-31, Mar. 1973.
- Alberto Sangiovanni-Vincentelli, D. Mandrioli, and Mauro Somalvico, A General Approach to Learning in Problem Solving, Computer Learning Processes, J. C. Simon, Editor, Nordhoff Co., Publisher pp. 471-501, 1976.

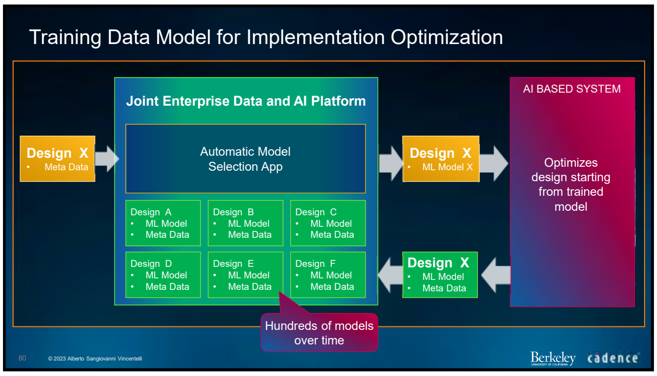
Berkeley cadence

78



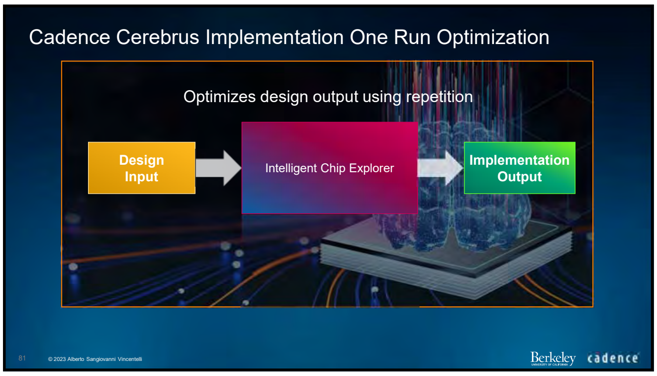
Opportunities for ML in Simulation
Regression compression
Bins Covered
CPU Time
393226
10052 cpuH
Original Regression
Bins Covered
CPU Time
Regain
Compression
390528
1950 cpuH
99.3%
5.1x
ML Regression
Targeted Regression
Original regression for full design
Augmenting runs from ML
Bug Hunting / Coverage Closure
prerequisite
cover properties
prop_1 25%
prop_2 60%
prop_3 5%
1%
prop_4 25%
0%
Original Regression
prop_1 35%
prop_2 60%
prop_3 15%
5%
prop_4 40%
2%
ML Regression focused runs
79
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

79



Training Data Model for Implementation Optimization
Joint Enterprise Data and AI Platform
Automatic Model
Selection App
Design A
ML Model
Meta Data
Design B
ML Model
Meta Data
Design C
ML Model
Meta Data
Design D
ML Model
Meta Data
Design E
ML Model
Meta Data
Design F
ML Model
Meta Data
Design X
Meta Data
Design X
ML Model X
Design X
ML Model
Meta Data
AI BASED SYSTEM
Optimizes
design starting
from trained
model
Hundreds of models
over time
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

80



Cadence Cerebrus Implementation One Run Optimization
Optimizes design output using repetition
Design
Input
Intelligent Chip Explorer
Implementation
Output
© 2023 Alberto Sangiovanni Vincentelli
Berkeley
cadence

81