The Phoenix and the future of computing systems design

Giovanni De Micheli

EPFL

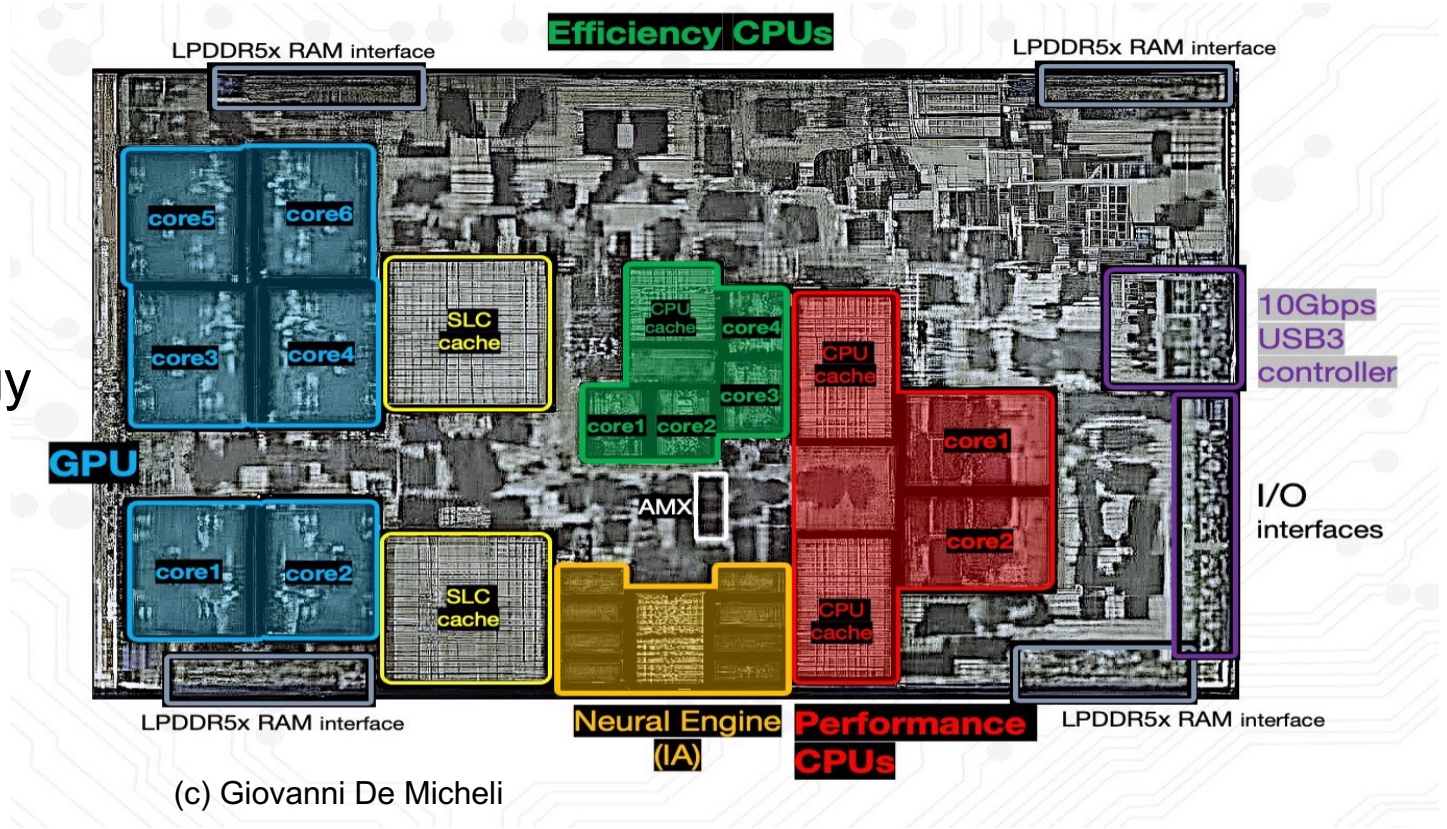
HiPEAC

LSI Integrated Systems Laboratory


Lausanne - CIO

The Systems on Chip motto

- Citius – Faster
  - Higher performance
- Altius – Higher
  - 3-Dimensional
- Fortius - Stronger
  - Do effort with less energy



The challenge for this decade and beyond

- AI/ML and data-driven science require:
  - Large storage space
  - Large amount of energy for computation
- How do we rethink architectures, circuits and devices ?
  - To enable *edge devices* to use AI/ML
  - To curb the energy comsumption?
- Human/planetary dimension of AI/ML

The future of computing system design

- Design is a cyclical process
  - Design closure is achieved by looping through synthesis and analysis tools
- What about when there is no closure ?
- Emerging technologies and novel circuits/architectures are required to go beyond the current energy and performance envelope
  - Reinvent design in different directions according to applications
- Design automation is the enabler for the rebirth of computing

The Phoenix

The Phoenix symbolizes the design of new systems
With new technologies and operation paradigms
Through the use of current digital design methods

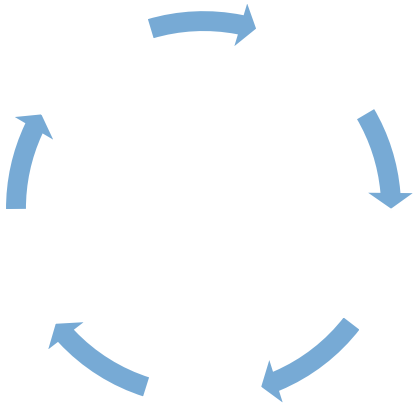
Looking beyond technology

- Advanced design and manufacturing requires a plurality of efforts and geostrategic components:
  - Most advanced design is realized in the West
  - Advanced fabrication is done in the Far East
  - EUV lithography requires machines built in Europe

- The future of computing systems depends also on global politics and economics

Peace and Happiness through Prosperity

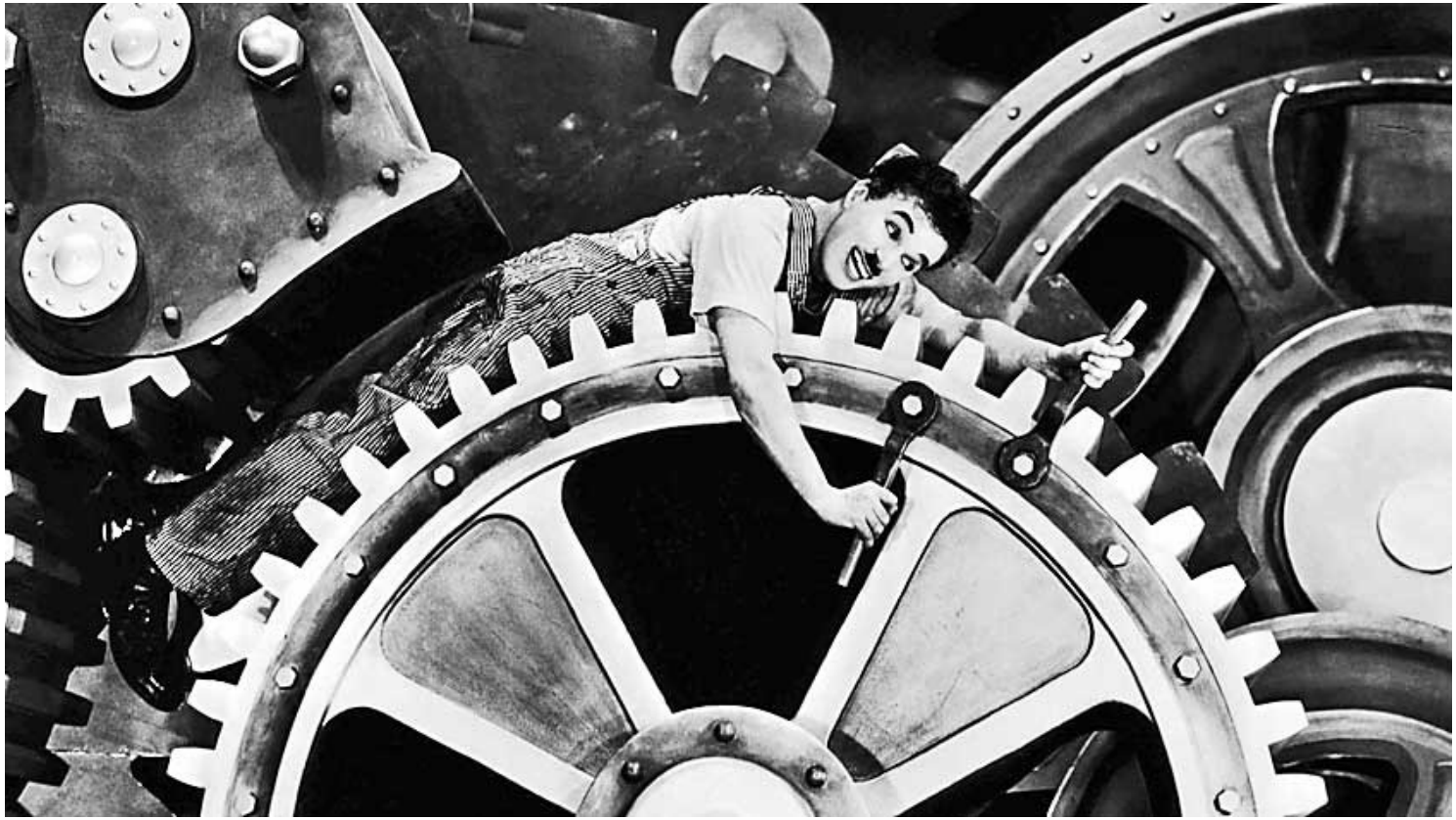
- Konosuke Matsushita (1894-1989)
  - Founder of Panasonic
- Peace and Progress through Prosperity
  - The three concepts are strongly related
  - Circular dependencies in PPP



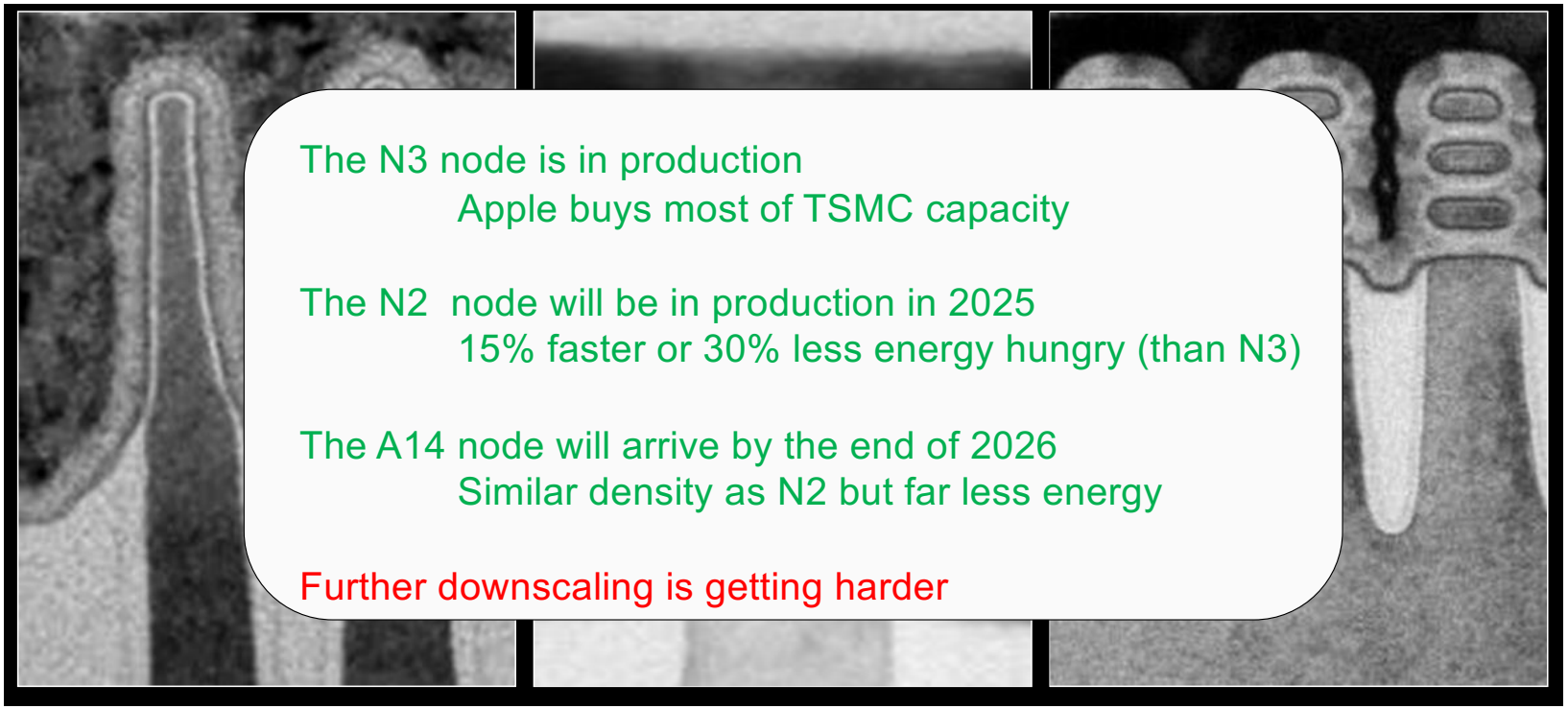
Key Point # 1

- We are at an inflection point:
  - Any major improvement in computing system requires a big step
- Peace and cooperation are essential to:
  - Pool exceptional talent and fabrication technology
  - Create broader markets supporting advanced manufacturing
- Funding computing research generously is crucial
- Supporting engineers with good rewards is key to retain them

The inside circles

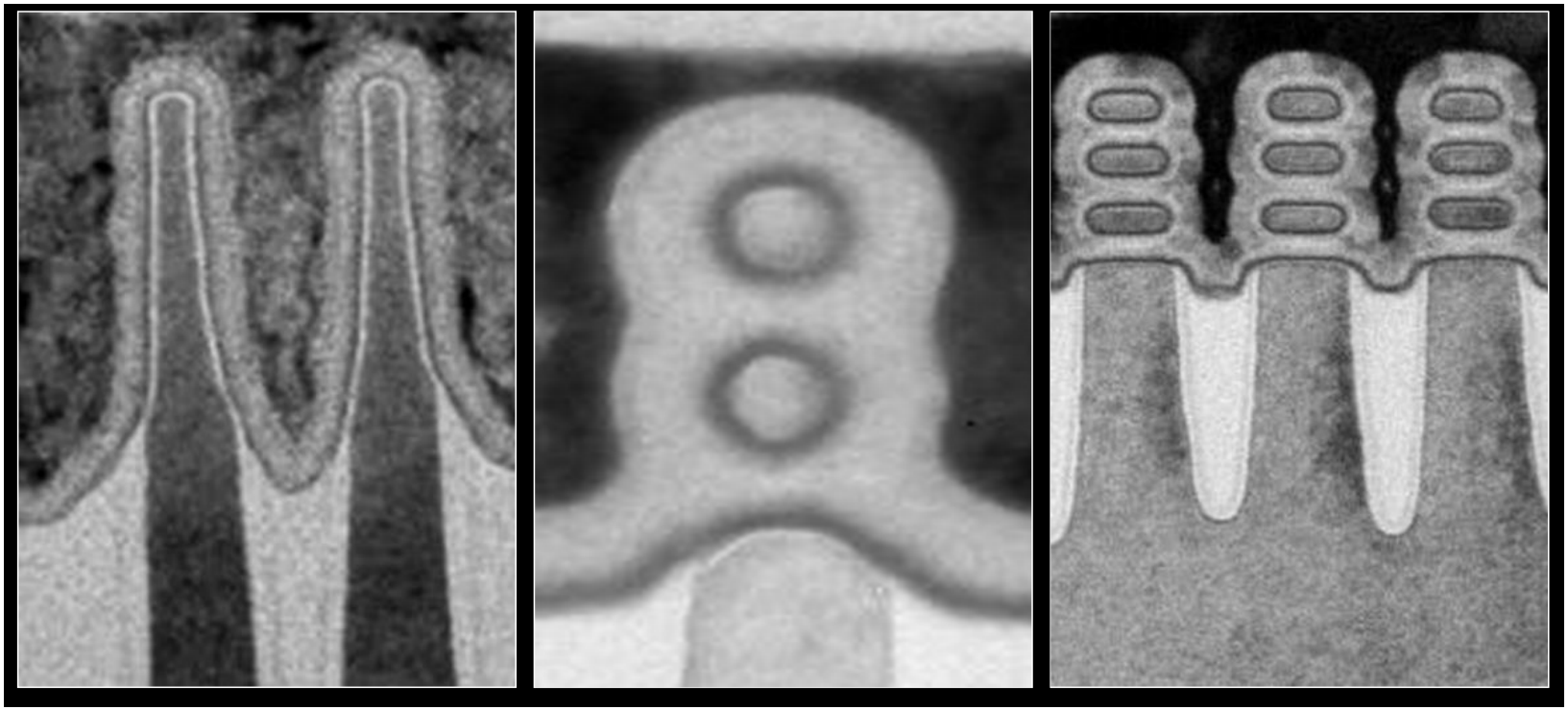


CMOS today: FINFETs, NanoWires and NanoSheets



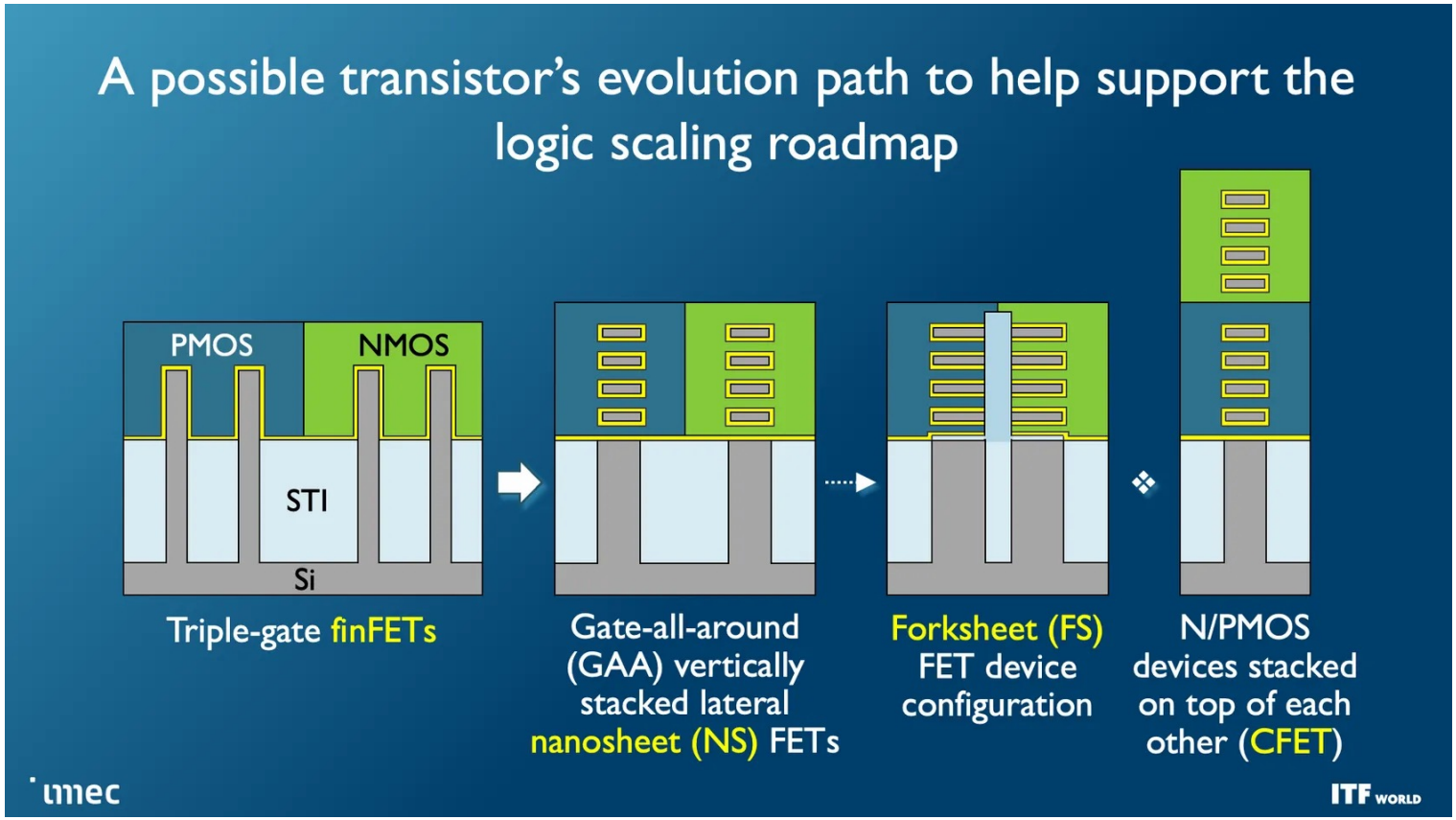
[INTEL, 2017]

CMOS today: FINFETs, NanoWires and NanoSheets



[INTEL, 2017]

CMOS tomorrow: Forksheets

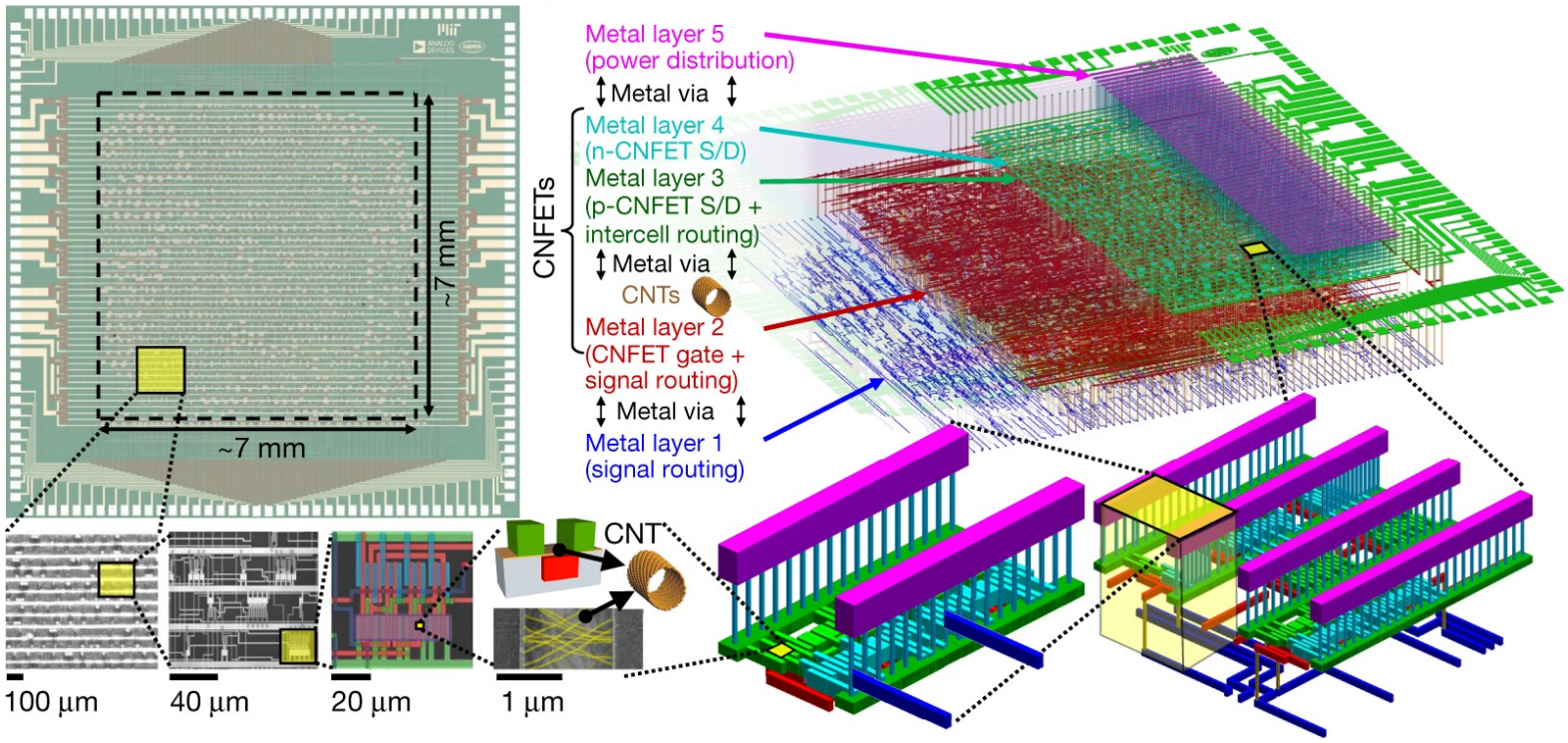


[IMEC, 2023]

Never ending scaling?

[IMEC 2023]

Will post-CMOS tecnologies be **always** the technologies of the future?

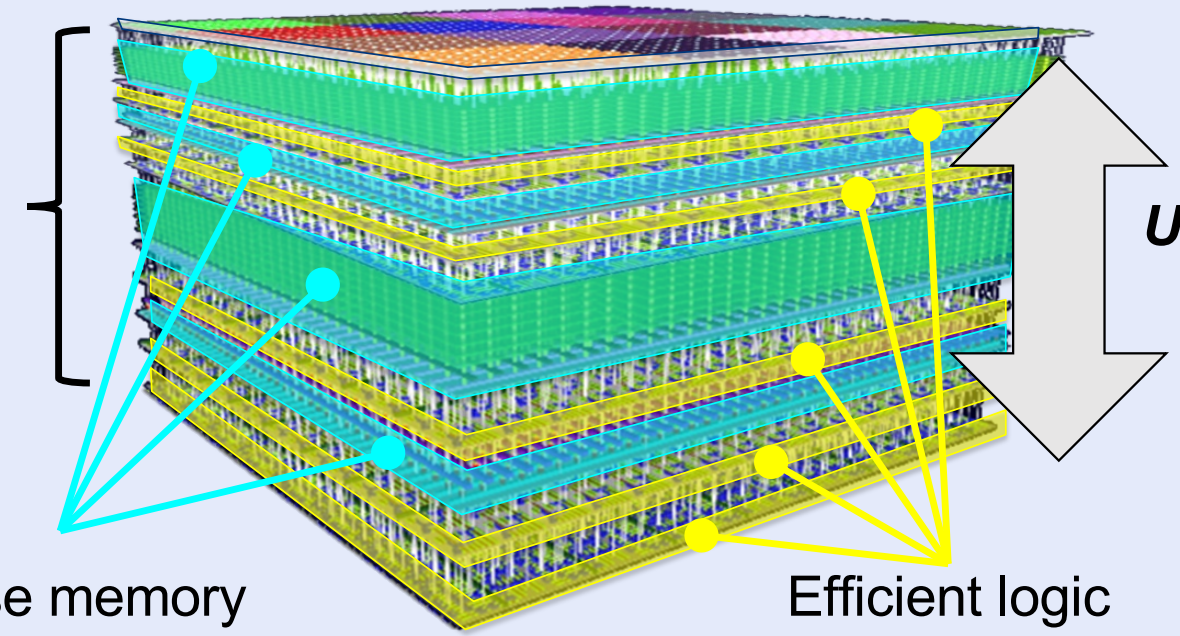


[Hills, Nature 19]

The N3EXT project

[Courtesy of S. Mitra, Stanford]

Interleaved logic & memory layers



Ultra-dense 3D

Dense memory

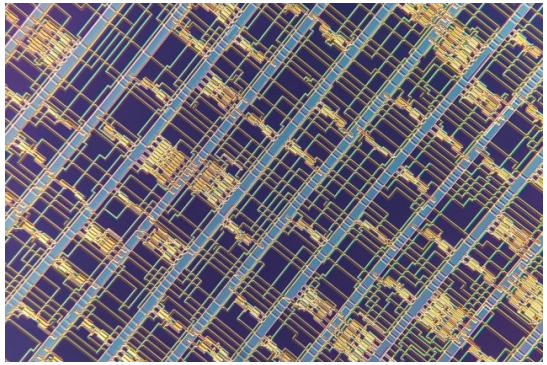
Efficient logic

100× – 1,000× Energy Delay Product (EDP) benefits

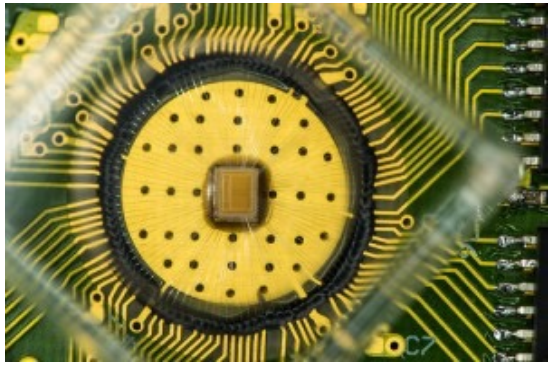
Diversity

Diversity

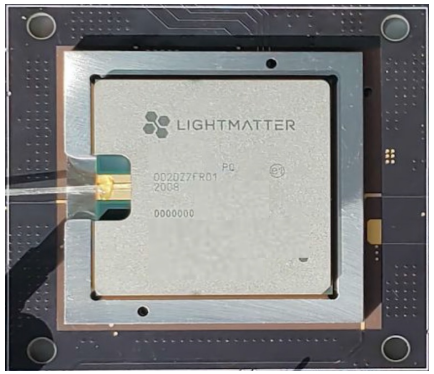
A plurality of technologies may be combined to achieve **acceleration** of computation and communication



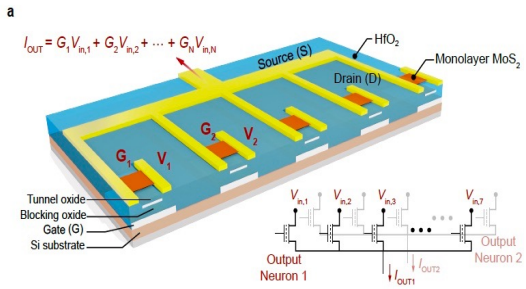
[Shulaker, MIT 2021]

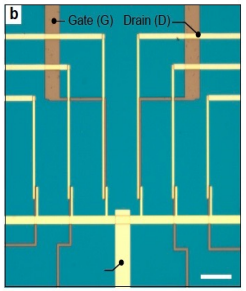


[Boybat, IBM 2020]



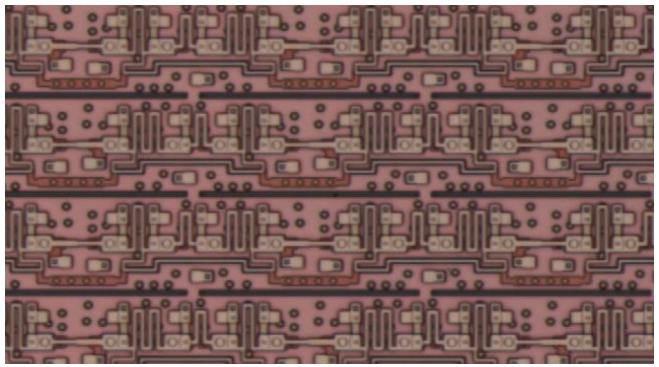
[Ramey, Hot Chips, 2020]





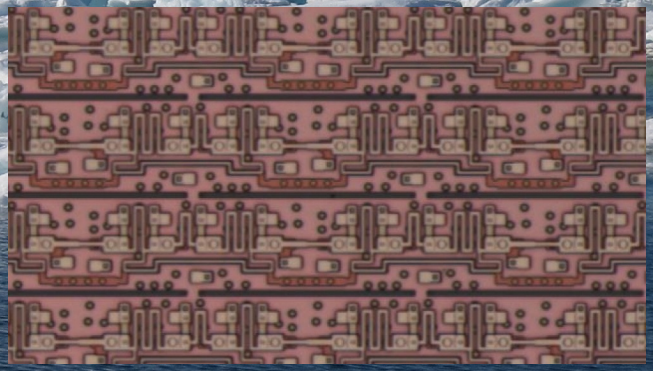
[Marega et al. ACS Nano 2022]

(c) Giovanni De Micheli



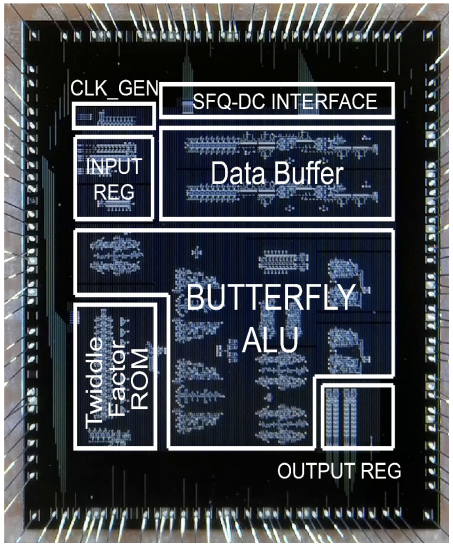
[RSFQ, Lincoln Labs, 2020]

The Cryo-world

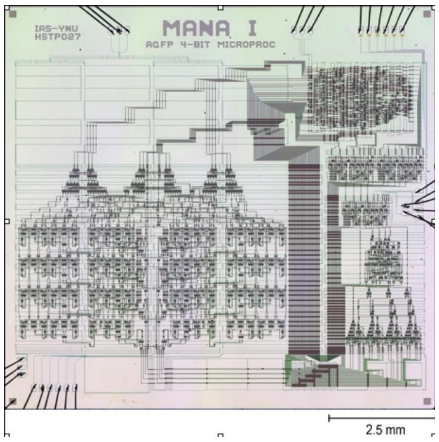


Superconducting electronics

- Technology
  - No parasitic resistance at low temperature (4K)
  - Information by *quantized pulses* $\int V(t)dt = \varphi_0 = h/2e = 2.07$ mv ps
- Design features
  - *Classic computing paradigm* with *deep pipelined* logic
  - Ultrafast computation with small energy consumption
  - Many variants including adiabatic operation
- *Majority logic* is the native abstraction in some realizations



47 GHz SFQ FFT Processor [Ke et al.,2021]



GHz 4-bit RISC [Ayala et a2.5l.,2021]

Superconducting electronics

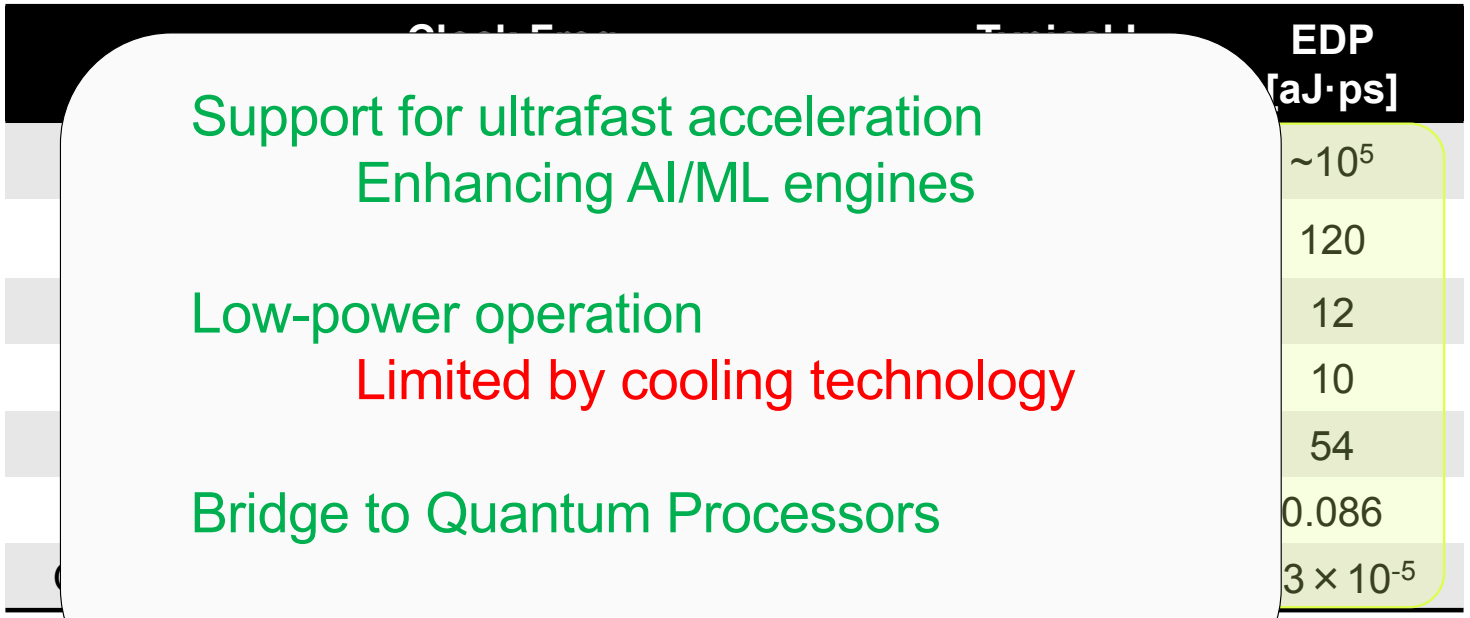
Logic	Clock Freq. [GHz]	$E_{bit} / I_c\Phi_0$	Typical I_c [mA]	EDP [aJ·ps]
CMOS	4	-	-	~10^5
RSFQ [1]	50	19	150	120
eSFQ [2]	20	0.8	150	12
RQL [3]	10	0.33	150	10
LV-RSFQ [4]	20	3.5	150	54
AQFP [5]	5	0.0083	50	0.086
Quantum limit	-	-	-	5.3×10^{-5}

Single-flux

Adiabatic

[1] X. Peng et al., IEICE Trans. Electron. **E97.C**, 188 (2014).
[2] M. H. Volkmann et al., Supercond. Sci. Technol. **26**, 015002 (2013).
[3] Q. P. Herr et al., J. Appl. Phys. **109**, 103903 (2011).
[4] M. Tanaka et al., IEEE Trans. Appl. Supercond. **23**, 1701104 (2013).
[5] N. Takeuchi et al., Supercond. Sci. Technol. 28, 015003 (2015).

Superconducting electronics



[1] X. Peng et al., IEICE Trans. Electron. E97.C, 168 (2014).
[2] M. H. Volkmann et al., Supercond. Sci. Technol. **26**, 015002 (2013).
[3] Q. P. Herr et al., J. Appl. Phys. **109**, 103903 (2011).
[4] M. Tanaka et al., IEEE Trans. Appl. Supercond. **23**, 1701104 (2013).
[5] N. Takeuchi et al., Supercond. Sci. Technol. 28, 015003 (2015).

Changing the rules

[P. Picasso, 1907]

Quantum computing

- Two facets of quantum computation:
- The physical world
  - Using matter to compute with *superposition* and *entanglement*
- The computational world
  - Designing and *compiling* algorithms into quantum circuits

[P. Picasso]

Quantum computing

Quantum computing

- Complexity of SW parallels complexity of HW
  - Parallel algorithm development
  - Quantum compilation into reversible circuits
- Error tolerance and corrections
  - Evolution from NIQC to FTQC
  - Relevant fields of applications
- QC will be a success in a restricted domain
  - Most problems will not fit in the QC paradigm

Thermodynamics of computation

- The search for efficient computing systems involves:
  - Reversible logic in quantum computing
  - Adiabatic computation is some superconducting families
- Thermodynamics of computing systems will be the limiting factor
  - Reduce/recycle energy cost and heat dissipation of computation
- Some circular models are researched
  - Use heat generated in water-cooled computers to produce electricty

EPFL

HEATING BITS

Key Point # 2

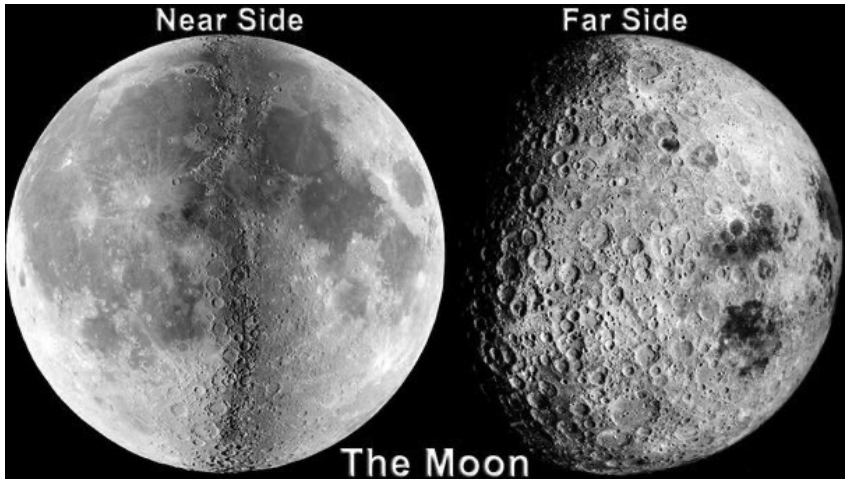
- The competitive advantage of emerging technologies:
  - Adapting computational model to computational fabric
- Application-specific accelerators are the best example
  - Very desirable for both cloud and edge systems
- What is the effort to achieve successful diversity?
  - New computational thinking models
  - Adapting/creating EDA tools and flows

Modeling the common traits of computing systems

- Search for common models
- Exploit expertise on contemporary logc design in new technologies and paradigms

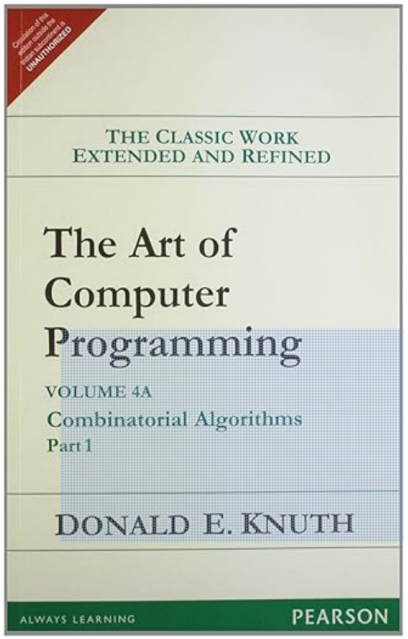
The basics and beyond

- We know a lot about logic networks and synthesis
- But there is more that we *don't know* than we *know*
- Is that relevant?



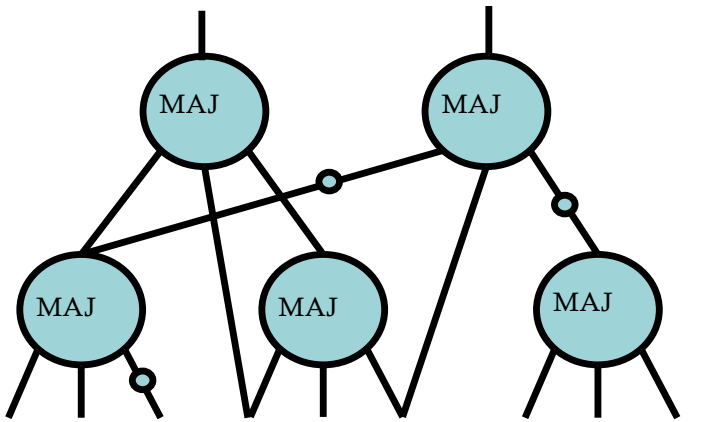
Logic Networks

- Interconnection of logic primitives
  - Simplest model: primitives are 2-input gates
  - Complexity of the network: # gates
  - Intractable problem
- Data structures for logic networks
  - Networks of NAND2 and/or NOR2
  - AIG – AND-INVERTER network: ABC [Mishchenko]
  - Many other models



Majority logic synthesis

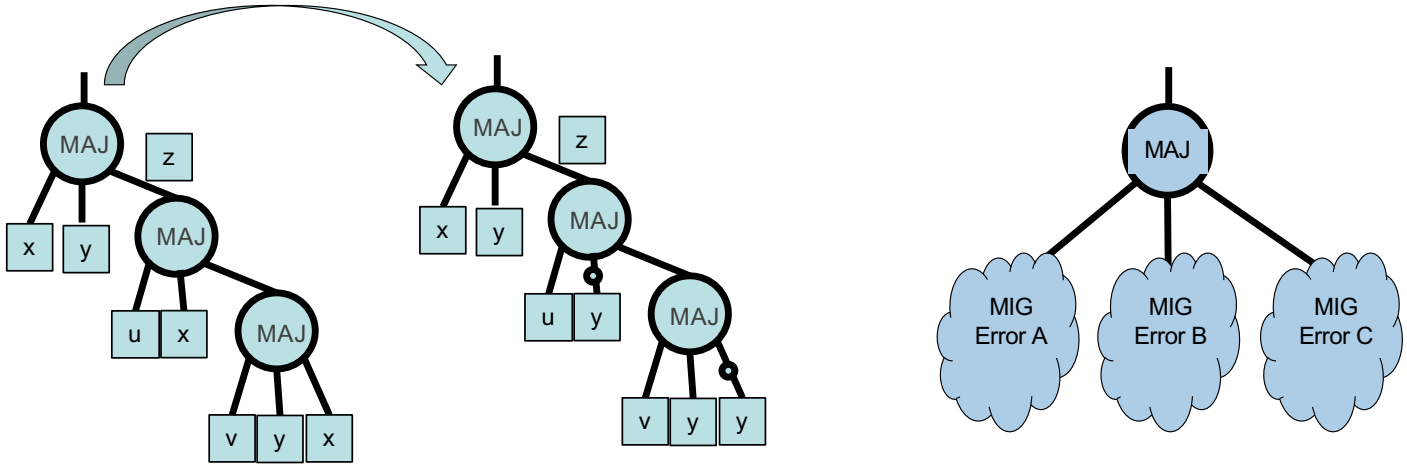
- Majority logic axioms
  - Sound and complete
  - Reachability property



[Amaru et al., DAC 14, TCAD18]

Algorithms for MIG optimization

- Optimization algorithms for area and delay
  - Algebraic methods based on rewriting
  - Boolean methods exploiting don't cares



Algorithms for MIG optimization

Majority-based algorithms:

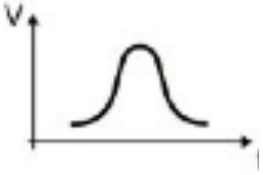
improve commercial and academic logic synthesis for **CMOS**
15-20% delay reduction as compared to earlier methods

are very successful for many **emerging technologies**, like:

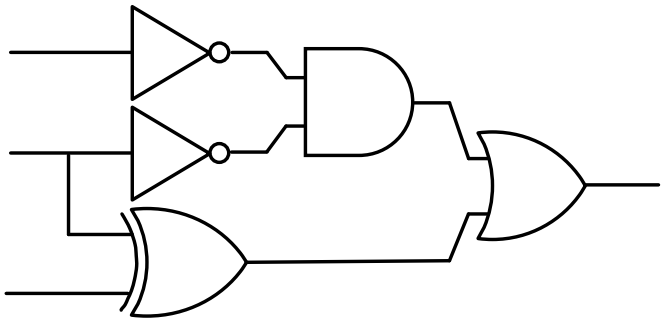
superconducting electronics
optical/wave based computing
logic in memory

Logic Synthesis for Superconducting Electronics

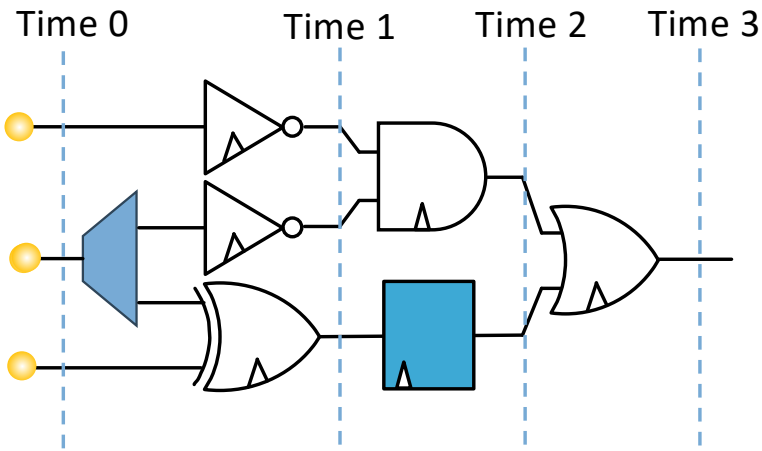
1. Limited driving capacity → adding splitters
2. Path balancing → adding delay elements



Classic CMOS



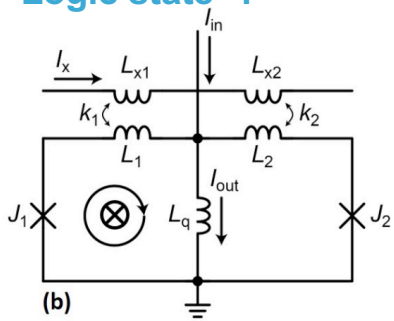
SFQ



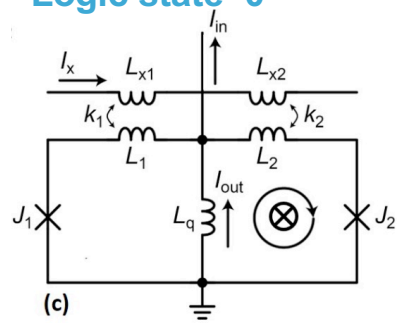
Adiabatic superconducting circuits (AQFP)

- *Adiabatic Quantum Flux Parametron*]
  - Adiabatic -> very small dynamic power consumption
  - Majority-based logic – Inversion available
  - MIG is the native data structure
- Logic optimization involves
  - Applying algebraic/Boolean optimization
  - Insert splitters and buffers

Logic state '1'



Logic state '0'



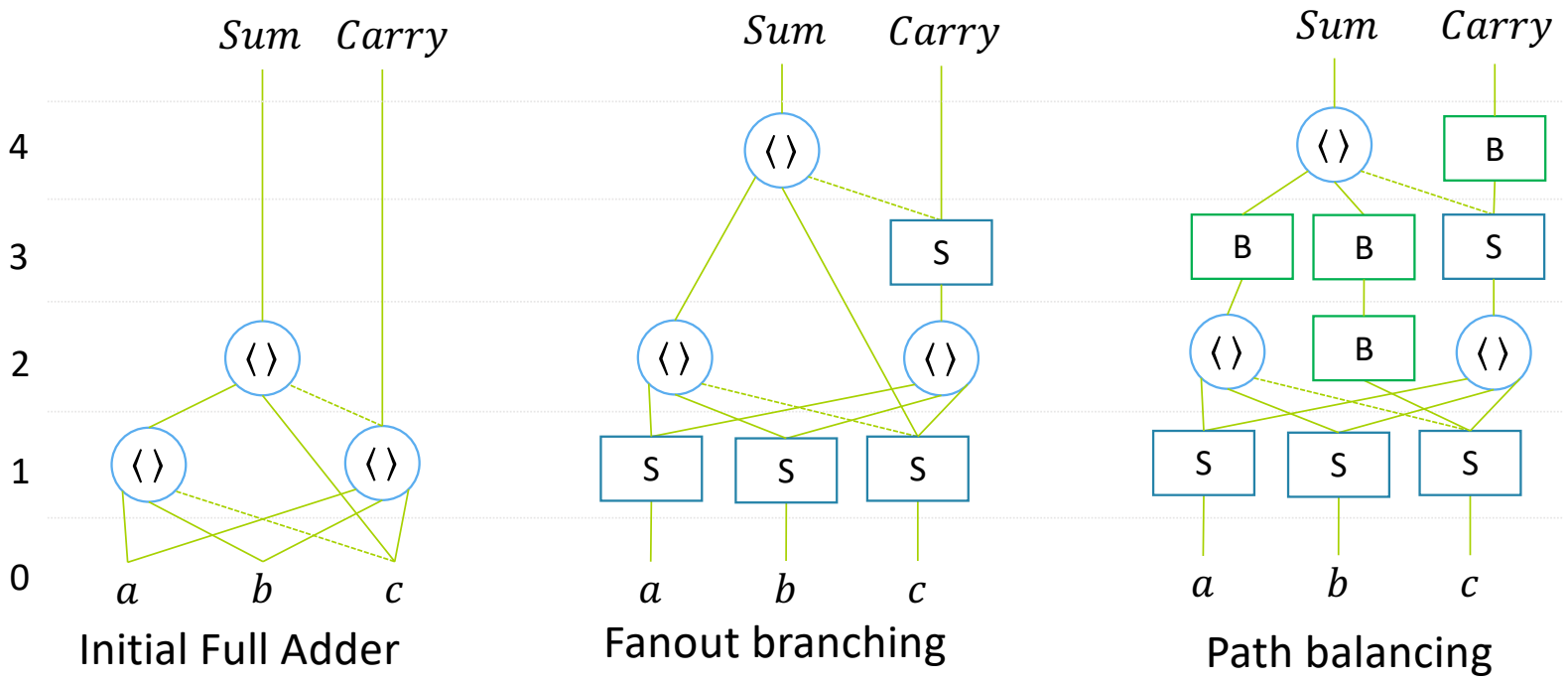
N. Takeuchi et al., "An adiabatic quantum flux parametron as an ultra-low-power logic device". Sup. Science and Tech. 2013

S.-Y. Lee and G. De Micheli. "Heuristic Logic Resynthesis Algorithms at the Core of Peephole Optimization", IEEE TCAD, 2023.

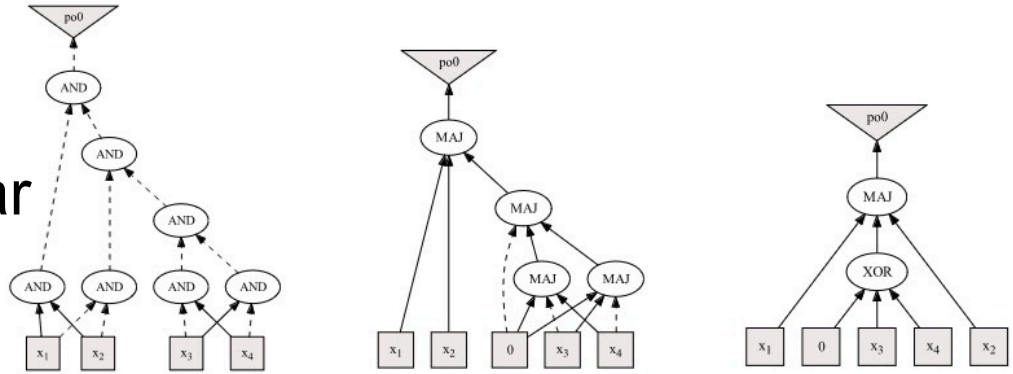
A. Tempia Calvino and G. De Micheli. "Scalable Logic Rewriting Using Don't Cares", DATE 2024, pp. 1–6.

Balancing and splitting in AQFP

- Splitters are clocked
  - Balancing and splitting are intertwined



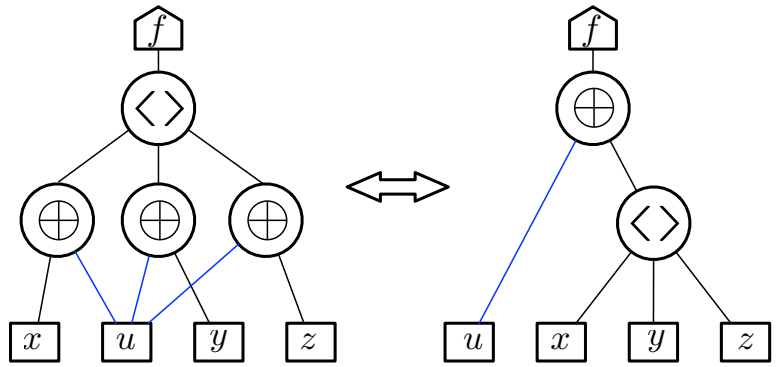
The role of linearity

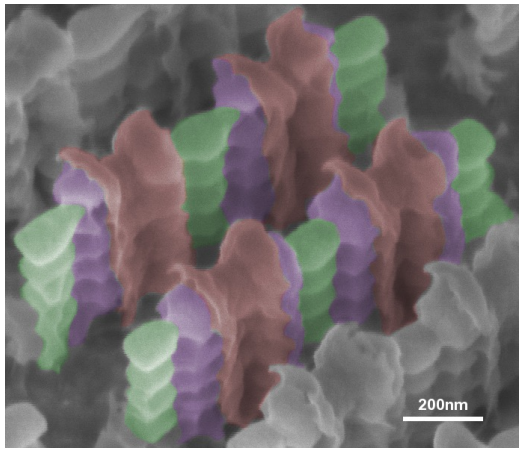


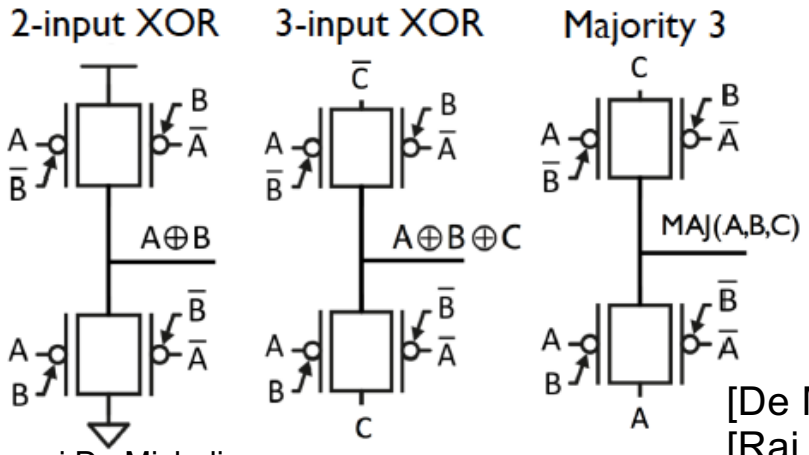
- AND-OR logic is inherently non-linear
- MAJority-Inverter logic as well
- EXOR are linear operators
  - EXOR play a leading role in realizing adders and multipliers
  - Adders require at least one AND gate per bit
- It is useful to decompose functions into linear and non-linear parts

XMGs – Controlled-polarity transistors

- XOR – MAJority graph
  - Extension of MIGs (replace INVerter by XOR)
  - Self-duality of XORs (odd) and Majority
- XOR can flow through MAJ nodes
- Data structure for controlled-polarity gates

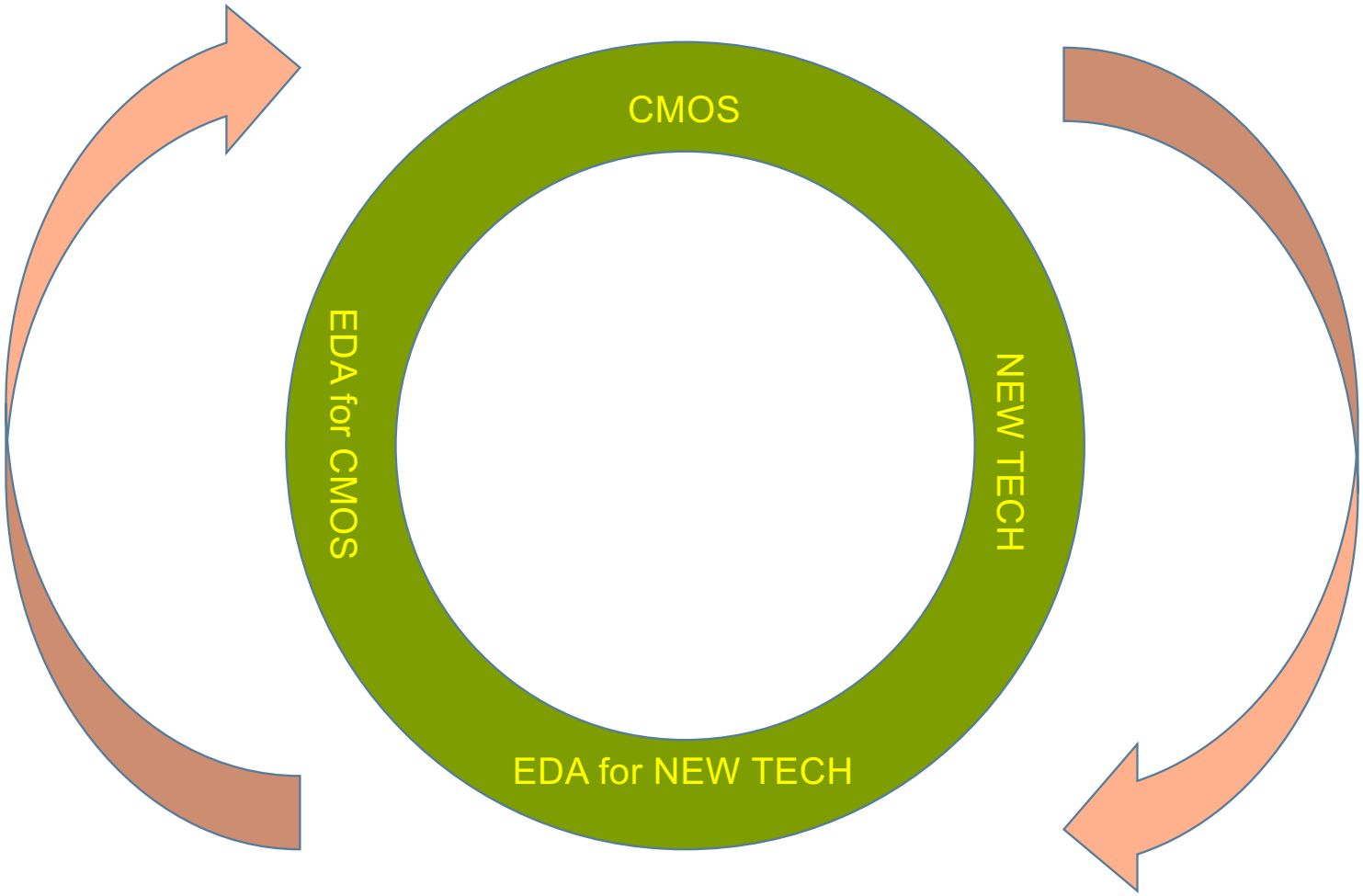






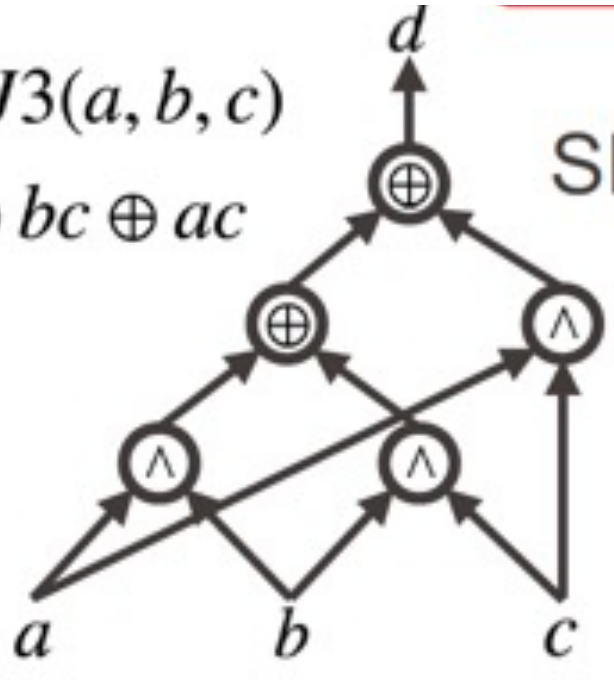
[De Marchi. IEDM 12]
[Rai et al., TCAD 22]

Innovation circle

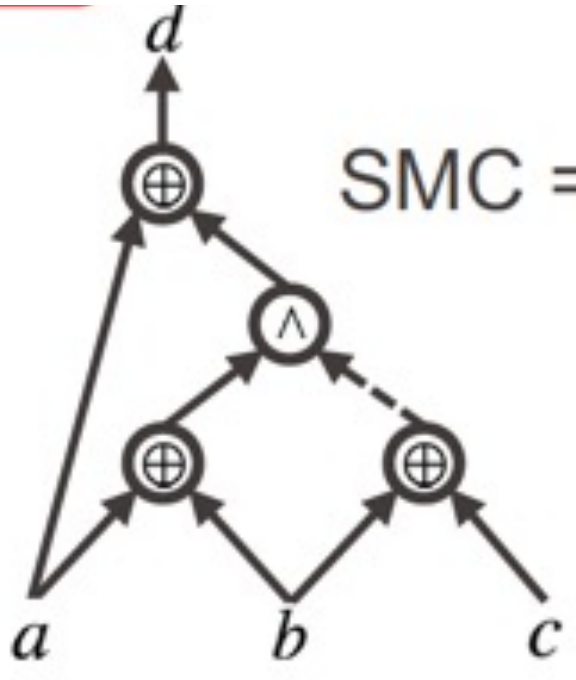


The role of multiplicative complexity

$$d = MAJ3(a, b, c)$$
$$= ab \oplus bc \oplus ac$$



SMC = 3



SMC = 1

security boil down to MC computation and MC-driven synthesis

Applications of XAGs - FHE

- Fully Homomorphic Encryption (FHE)
  - Compute functions using encrypted data
  - Speed of computation is essential
- Apply logic synthesis to restructure computation
  - While keeping noise under a threshold level
- Key metrics:
  - Multiplicative complexity – multipliers generate noise
  - Multiplicative depth – critical path
- Approaches:
  - MC-aware MD optimization (MC * MD^2)
  - ESOP balancing

[Yu, WAHC 24]

Applications of XAGs - QC

- Quantum compilation requires reversible logic synthesis
  - Mapping logic to library, e.g., Clifford+T library
  - Cost of primitives varies: T cells are most expensive
- Modeling quantum oracles with XAGs
  - Goal is optimizing multiplicative complexity
  - Multiplicative complexity correlates with T-count
- Similar considerations apply to multiplicative depth

[Meuli Nature QI, 22]

Key Point # 3

- Most data flow computations can be abstracted by graph models
- Logic synthesis enables various optimization of graph models
- Some approaches and goals are shared across technologies
  - Critical path minimization
  - Complexity minimization
- Multiplicative complexity reduction is key to solving various problems

Summary and conclusions

- Progress requires stability and international collaboration
  - Especially important for the evolution of computing systems
- Efficient computing needs a match between architecture and technology
  - Accelerators will exploit the diversity of materials and circuits
- Logic modeling and reasoning is a fundamental tool
  - Algorithms for diverse design may leverage a wealth of techniques
- Chip design is entering a new exciting epoch
  - To sustain the fast evolution of algorithms, software and services

Thank You

Thank You