Superconducting Digital Computing: Promise-Progress-Prospects

Quentin Herr, Anna Herr

Semicon Europa, Messe München

18 November 2021

"These then are some illustrations of things that are happening in modern times—the transistor, the laser, and now these [Josephson] junctions, whose ultimate practical applications are still unknown."
— Richard Feynman, 1965

Key Industry Challenges

Industry dynamics

- Cost and environmental impacts of AI are unsustainable
- Hardware does not keep up with AI demands
- Static training at data centers is inefficient
- Next-node foundry costs are rising

Superconducting solutions

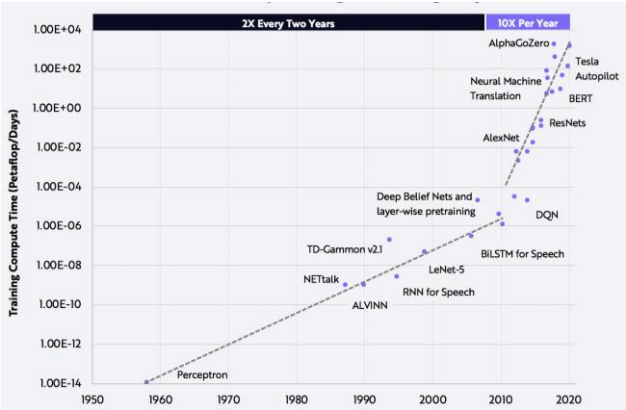
Energy efficiency

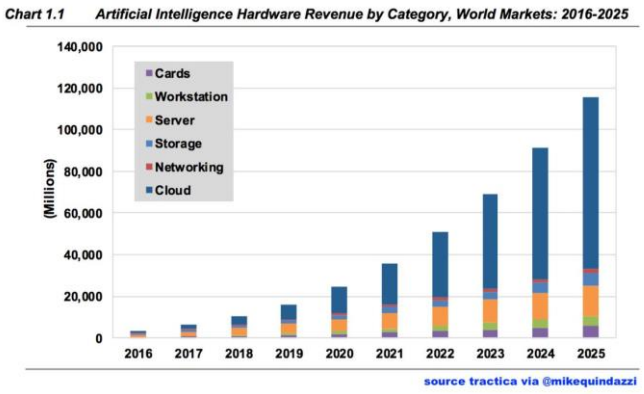
Unmatched compute density

Local, real-time systems

Quantum computing

Unlimited demand for compute at any cost





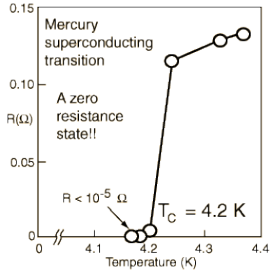
Source: ARK Investment Management LLC, "AI and Compute." OpenAI, https://arkinv.st/2ZOH2Rr.

imec

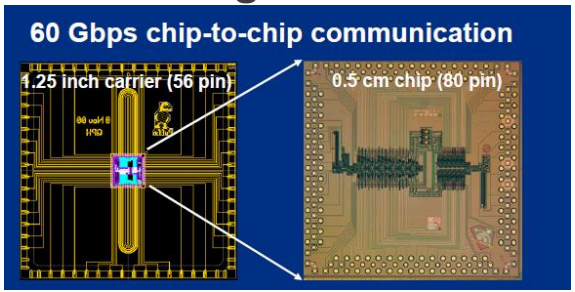
Promise

Fundamental advantages of superconducting digital electronics

Zero resistance wires



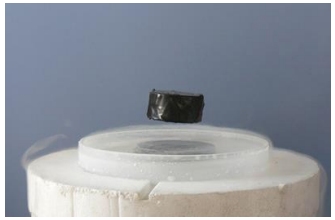
700 GHz Analog interconnects



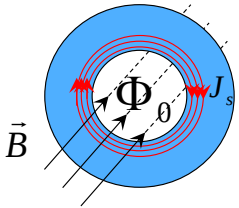
Herr, Quentin P., Andrew D. Smith, and Michael S. Wire. "High speed data link between digital superconductor chips." *Applied physics letters* 80, no. 17 (2002): 3210-3212.

A. Y. Herr et al., "An 8-bit carry look-ahead adder with 150 ps latency and sub-microwatt power dissipation at 10 GHz," Journal of Applied Physics, vol. 113, no. 3, p. 033911, 2013.

Meissner effect

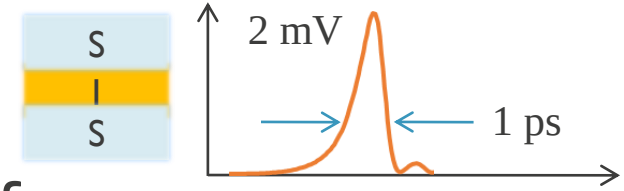


Quantum accurate digital bits



$$\Phi_0 = \frac{h}{2e} \approx 2 \times 10^{-15}\ \text{Wb} = 2\ \text{mApH}$$

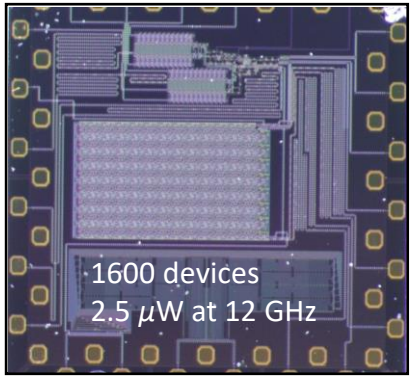
Josephson effect



Fast, low energy digital logic

$E = 2\ 10^{-20}$ J

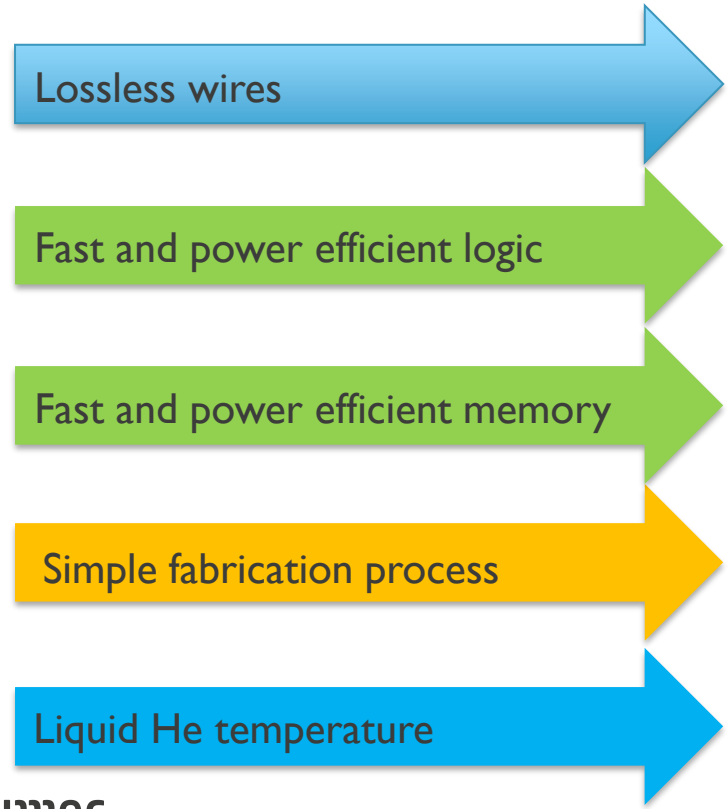
$\Phi_0 \approx 2$ mVps



imec

Compute with Single-Flux-Quantum Logic

Fundamentals



Architecture advantages

- Unmatched interconnects bandwidth
- Negligible energy dissipation
- High throughput & latency
- Dense packaging
- System volume cooling
- Native co-processor for superconducting quantum computer

imec

Application and Market Space

- Large ML workloads
- ExaFlop in a single system
- Fast Terabit data analytics

- Supercomputer modules
- HPC work loads

- Classical/quantum co-processor
- Neuromorphic co-processor

EDGE

Cost effectiveness
Low latency
Local model effectiveness
Reinforcement learning
Fast, real-time processing

BIG DATA

Cost effectiveness
Energy efficiency
Increased throughput

QUANTUM

Changing computing paradigm
Enabling quantum algorithms

Classical-Quantum Integration

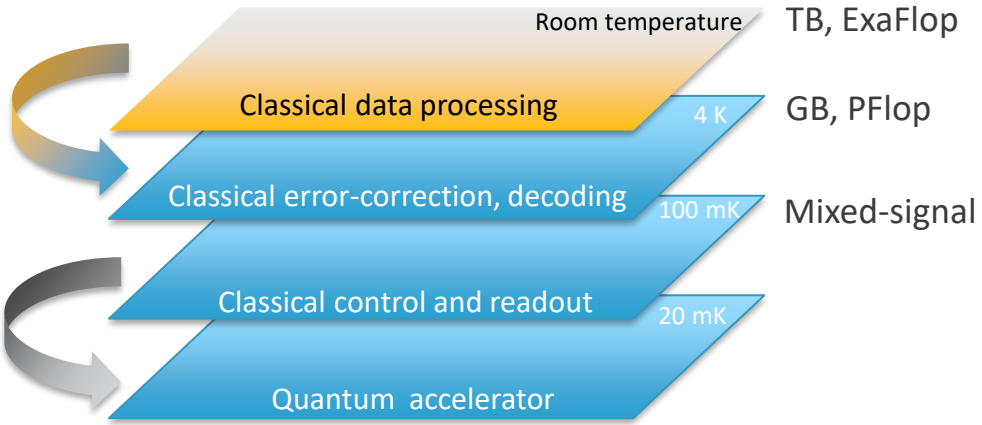
Superconducting technology key advantages

- Compatibility in operating temperature
- Compatibility in energy levels
- Compatibility in materials and devices

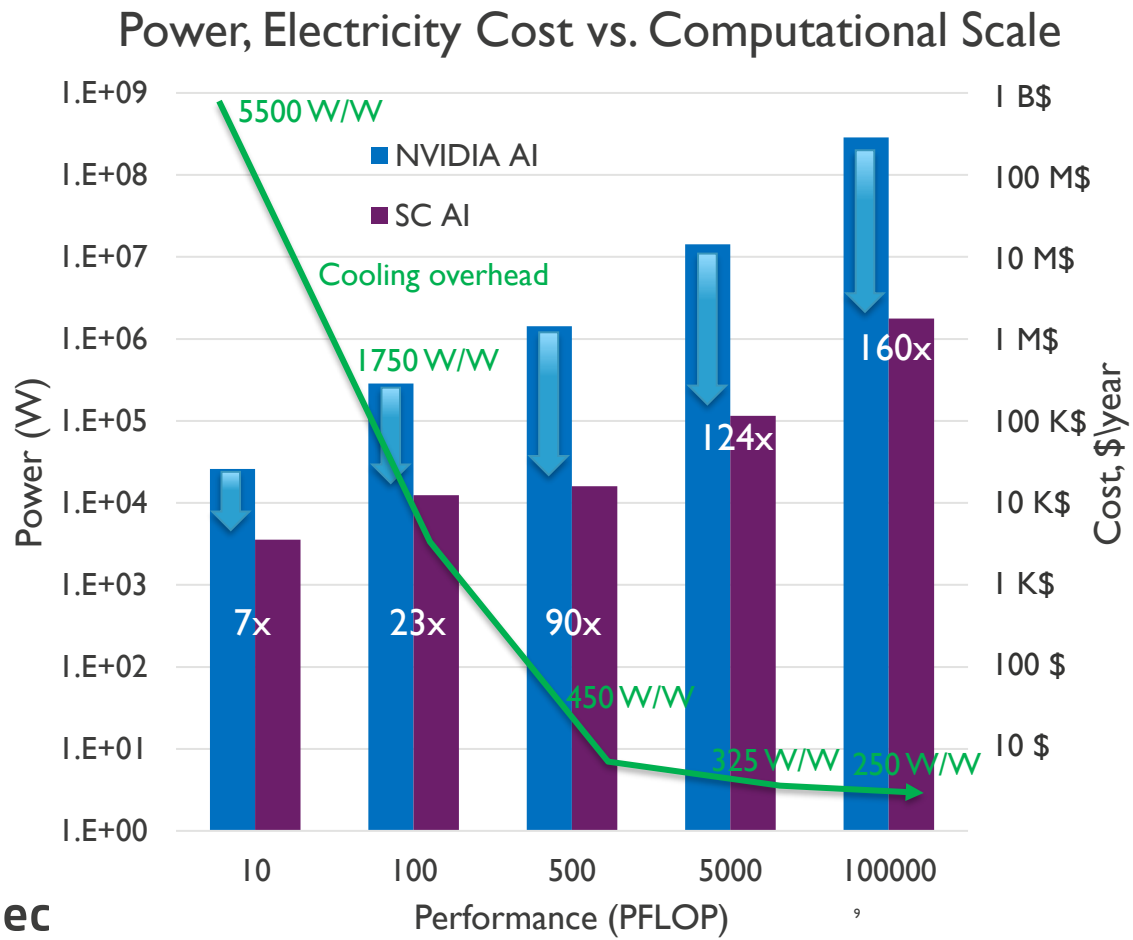
Reducing latency through the stack
Moving massive processing to 4K

Reducing energy disparity
Quantum control with minimal noise

Sharing the same fabrication process
Cost efficiency



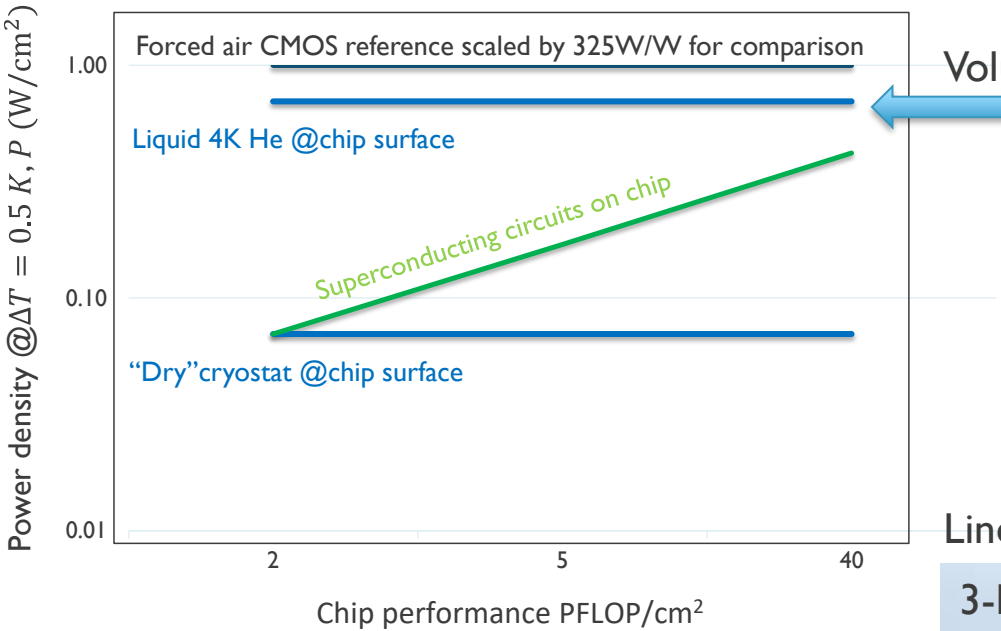
Efficient Liquid He Cooling at Large Scale



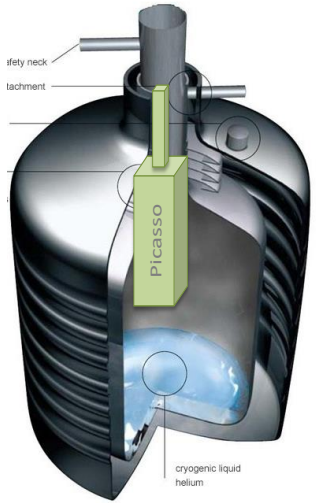
- Superconducting electronics breaks even at PFLOP scale
- Rapid increase in power efficiency with scale
- **100 M$/year** savings in electricity

imec

Volumetric Cooling Enables Dense Packaging



Volumetric He cooling



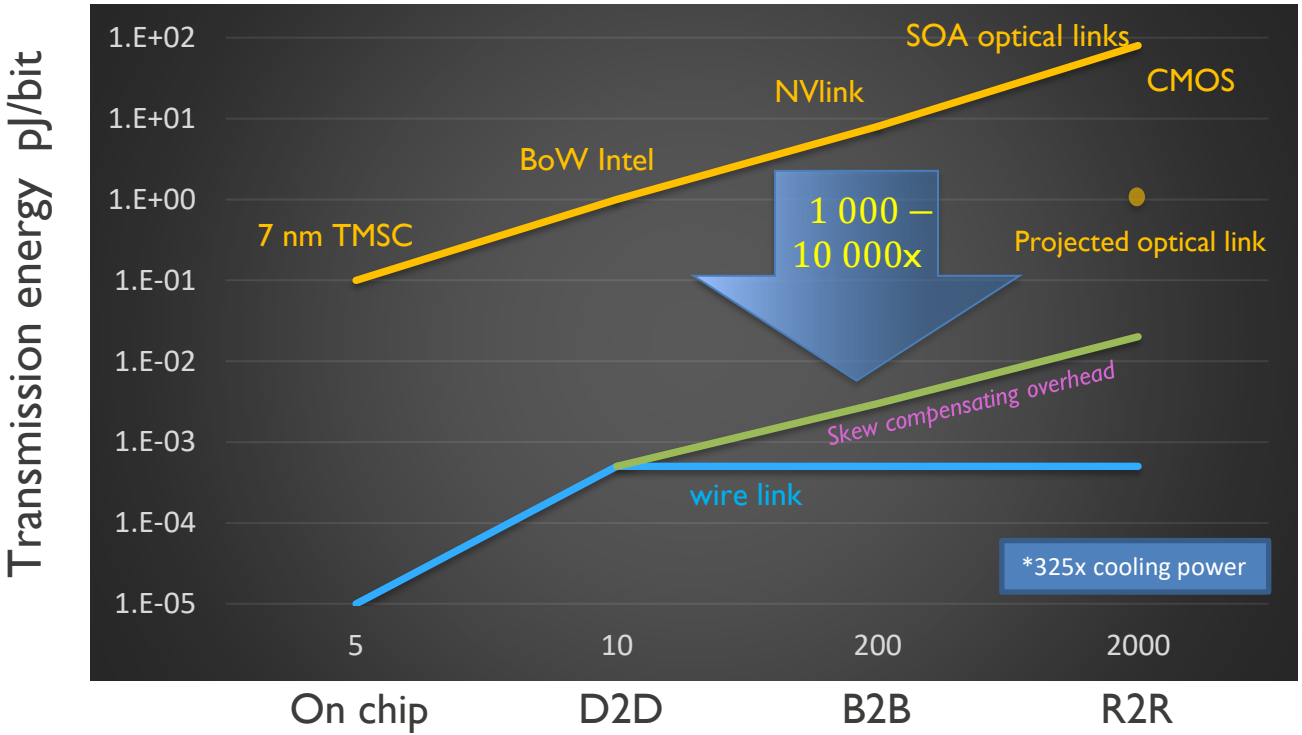
Linde commercial cryocooler

3-Rack form factor
1 KW cold/320 KW wall plug
1 ExaFlop compute power
2 M$ low volume production

imec

Interconnects with Breakthrough Energy Efficency

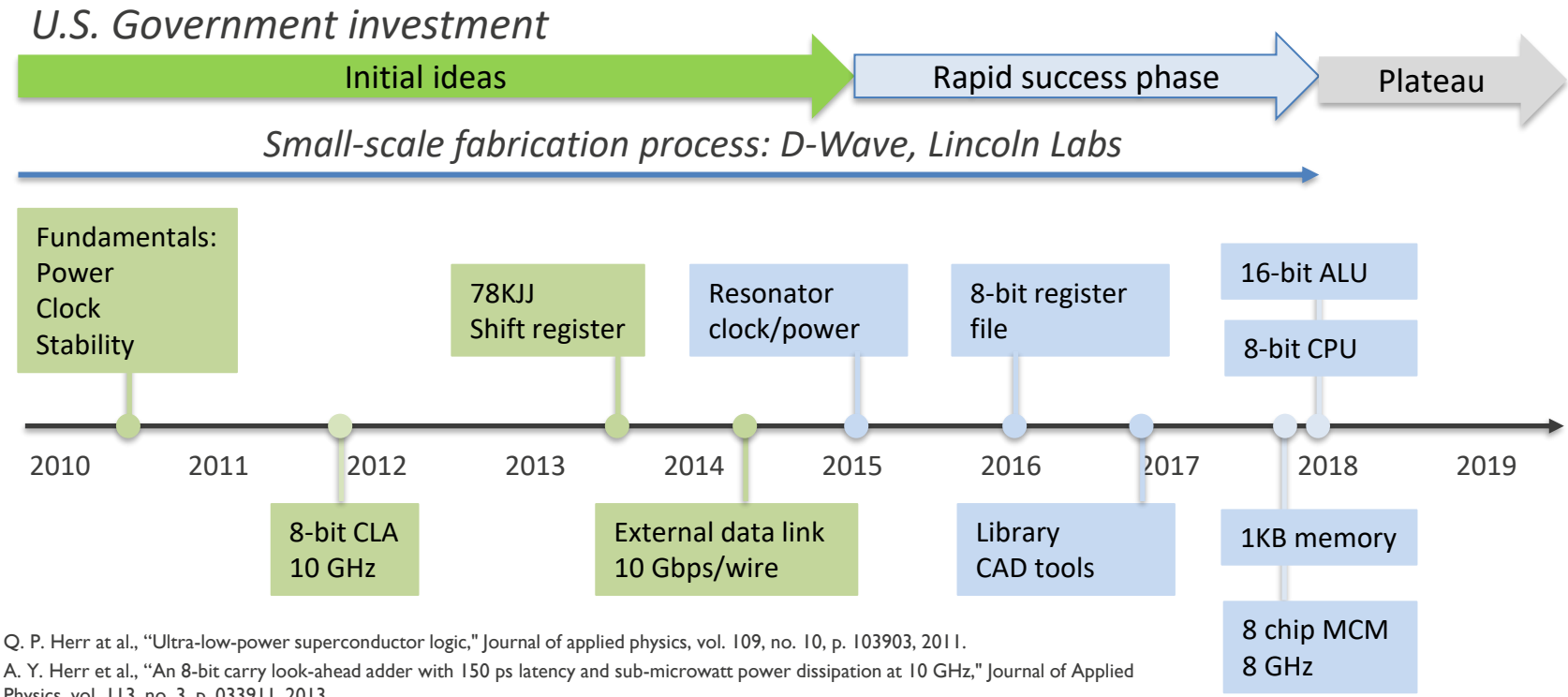
Superconducting Wires are Lossless from DC to 700 GHz Analog Bandwidth



- Losses 10^{-6} per wavelength
- Pulse based
- Source terminated
- 50-100 Ω
- Energy per bit 0.1 atto- J

Progress

Feasibility Stage of Superconducting Digital Development



Q. P. Herr at al., "Ultra-low-power superconductor logic," Journal of applied physics, vol. 109, no. 10, p. 103903, 2011.

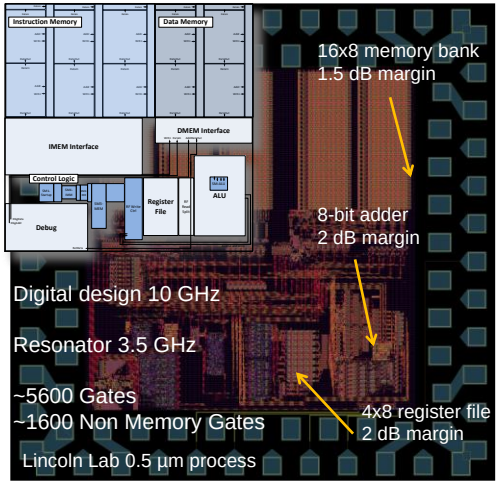
A. Y. Herr et al., "An 8-bit carry look-ahead adder with 150 ps latency and sub-microwatt power dissipation at 10 GHz," Journal of Applied Physics, vol. 113, no. 3, p. 033911, 2013.

Q. P. Herr et al., "Reproducible operating margins on a 72 800-device digital superconducting chip," Superconductor Science and Technology, vol. 28, no. 12, p. 124003, 2015.

H. Dai et al., "Isochronous data link across a superconducting Nb flex cable with 5 femtojoules per bit," arXiv preprint arXiv:2109.01808, 2021.

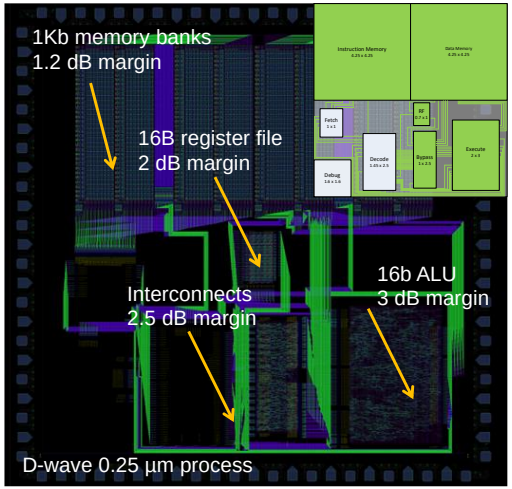
Core Technology Demonstrations

8-bit CPU 5x5 mm^2



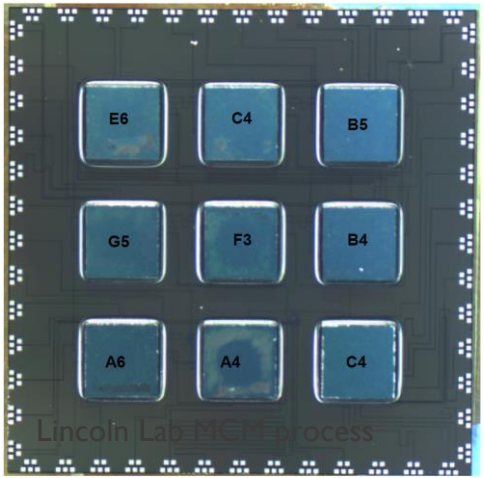
- Resonant clock/power
- RISC CPU architecture
- Gate library
- Memory access

16-bit CPU 10x10 mm^2



- Compatibility with CMOS RTL
- Efficient EDA tools
- Ultra fast interconnects
- Multibank memory

9 chip MCM 32x32 mm^2



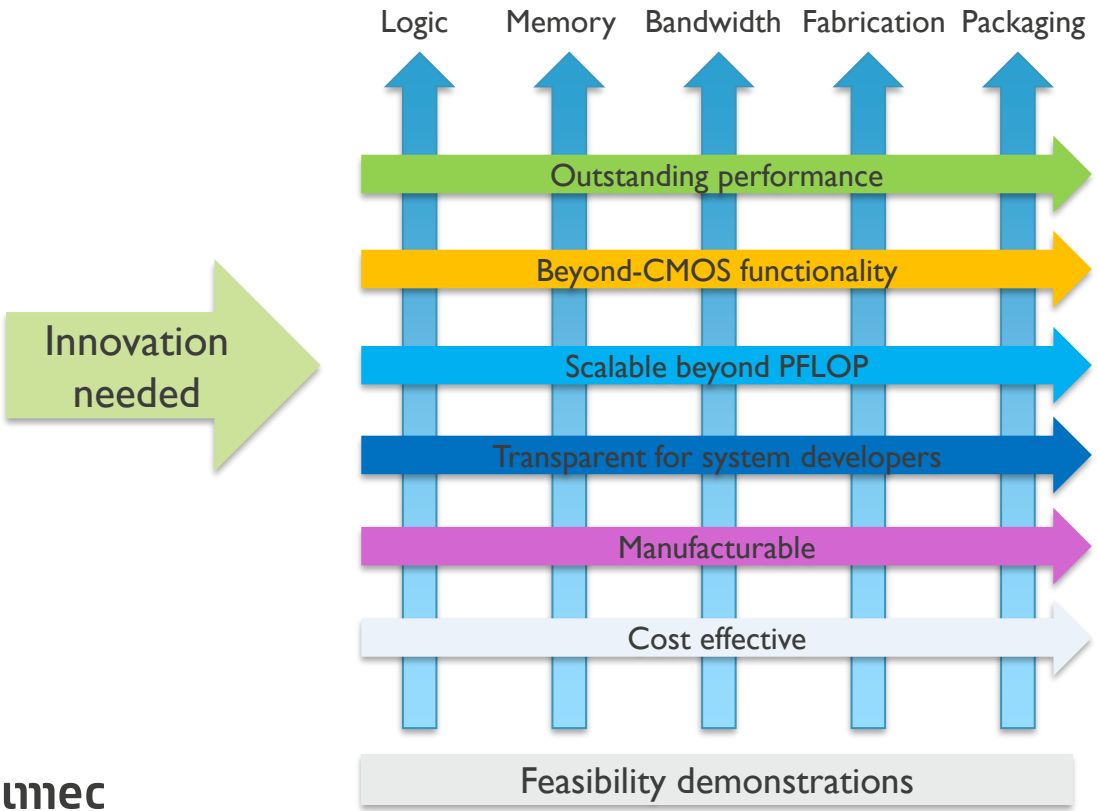
- Superconducting MCM
- Resonator scaling
- D2D communication at speed
- Synchronous communication

M. Vesely et al., “A pipelined superconducting 16-bit CPU design,” Presented at the Applied Superconductivity Conference, Washington State Convention Center, Seattle, WA, October 29, 2018.

J. Egan et al., “Synchronous chip-to-chip communication with a multi-chip resonator clock distribution network," arXiv preprint arXiv:2109.00560, 2021.

Superconducting Technology Enablers

A deficiency in any of a number of factors dooms an endeavor to failure (Moore, 2001)



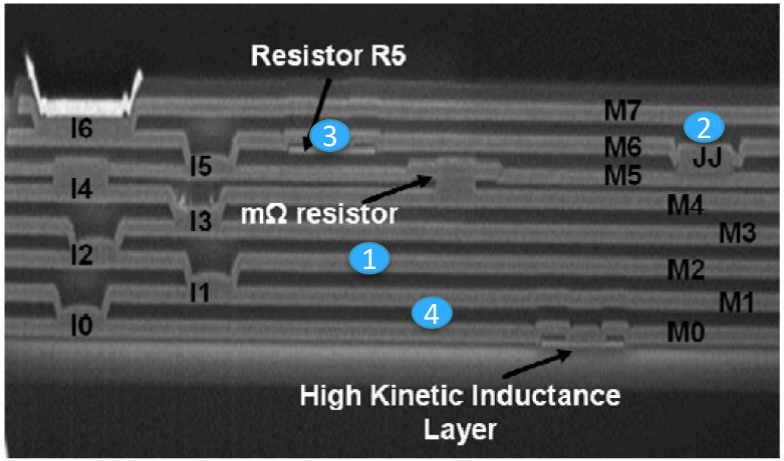
Window of opportunity

- Exploding world of information
- Demand for green hardware
- Change of compute paradigms

imec

Superconducting Fabrication Process

Lincoln Lab “SFQ5ee” process



General Features

1 MJJ

0.25 µm

4-8 metal layers

Step coverage vias

Sergey Tolpygo et al., “Advanced fabrication processes for superconducting very large-scale integrated circuits.” IEEE Transactions on Applied Superconductivity 26, no. 3 (2016): 1-10.

1. Nb wiring $T_c \approx 9$ K
   - Sputtering
   - Getter material
   - Temperature budget < 200° C
   - Refractory metal
   - Minimal feature size $\approx 0.25\ \mu m$
2. Nb/AlOx/Nb tunnel Josephson junctions
   - Sputtering & in chamber oxidation
   - Good wafer-to-wafer uniformity
   - Temperature budget < 150° C
   - Critical current density < $100\ \mu A/\mu m^2$
   - Thin barrier hard to control
   - Minimal feature size R $\approx 0.34\ \mu m$
3. Normal metal shunt
   - Required to shunt JJ capacitance
   - Sputtering
   - Large area
4. Low temperature TEOS ILD
   - High loss
   - Poor mechanical stability

Prospects

Switching Material Basis Enables Fabrication Process Advance

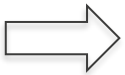
Leveraging progress in the Quantum and RF communities

Process modules

Josephson junction

High critical current density
Low capacitance
High quality factor
High thermal budget

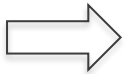
α-Si Barrier



Efficient wires

High adjustable inductance
High thermal budget
High temperature stability

NbTiN



Performance metrics

Density and Speed

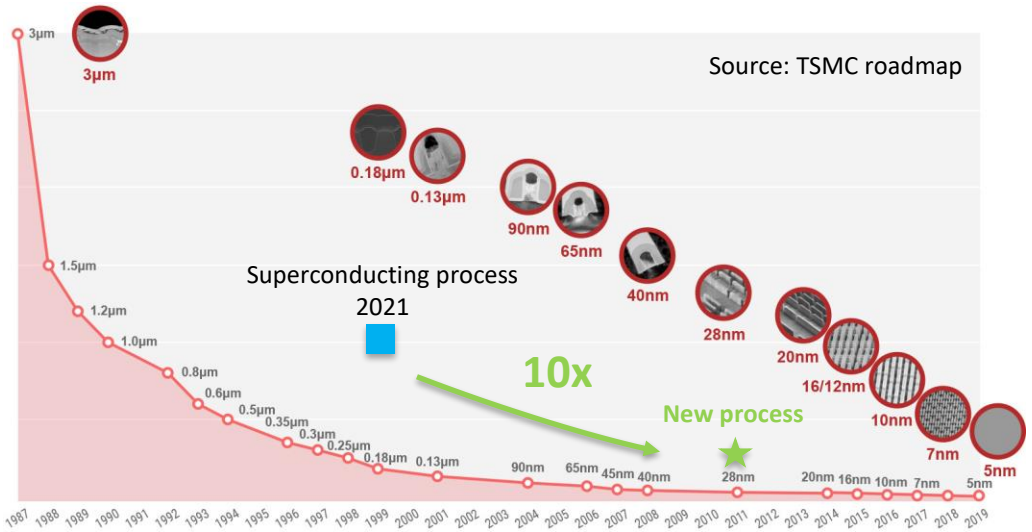
Minimum device diameter 210 nm
Clock frequency 30-50 GHz
Low spread < 2%
Easy integration

Density and Power Efficiency

The smallest pitch 100 nm
Power delivery efficiency 89%
Efficient stack with 16 ML for 0.4 BJJ/cm^2

imec

Scaling Superconducting Fabrication



20 ExaFlop system

- 0.4 BJJ/cm^2
- 16 ML stack
- 24-50 GHz clock frequency
- 200x power efficiency

New process start

10 years ahead in “superconducting” CMOS scaling

10 years behind CMOS scale

Material basis for scaling down to 10 nm

Architecture Trade Space

	CMOS 7 nm	Superconducting 30 nm
Speed	1.4 GHz	**17x**
Memory	500 MB/cm^2 (SRAM)	**0.01x**
Device density	1T devices/cm^2	**0.1x**
Interconnects	1.6 Gb/line @ 1 pJ/bit	**120x, 1000x**
Power efficiency	1 TOPS/W	**50x**

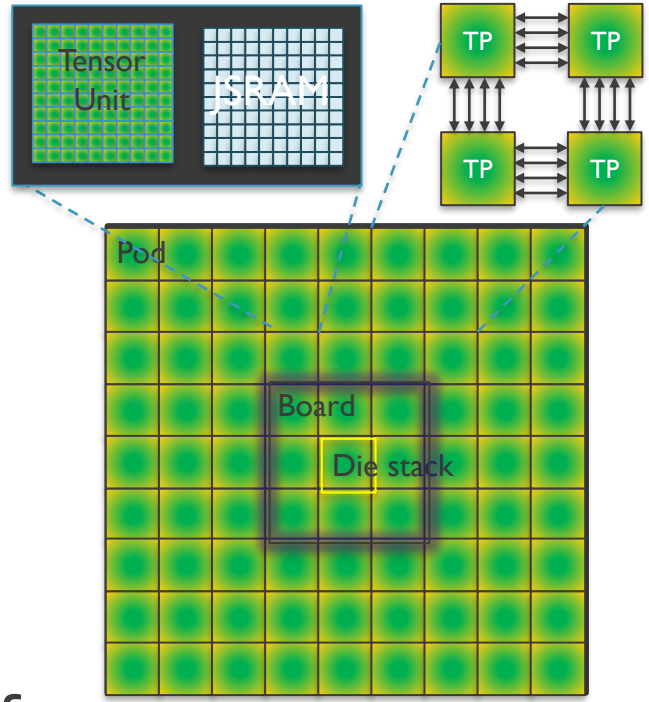
Trading density and memory capacity for speed and interconnect bandwidth

- Large work loads with high arithmetic intensity and memory reuse
- Dense packaging with extreme interconnect bandwidth

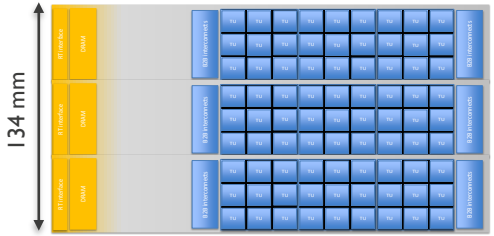
imec

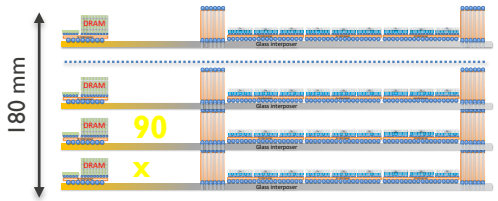
Packaging Optimized for Dense Workloads

Illustration: Flat GMM architecture
Sufficient bandwidth to fully interconnect



- 325 KW/ 1 KW (cold)
- 20 AI ExaFLOPS in 0.001 m^3
- 100x performance vs NVDIA DGX architecture



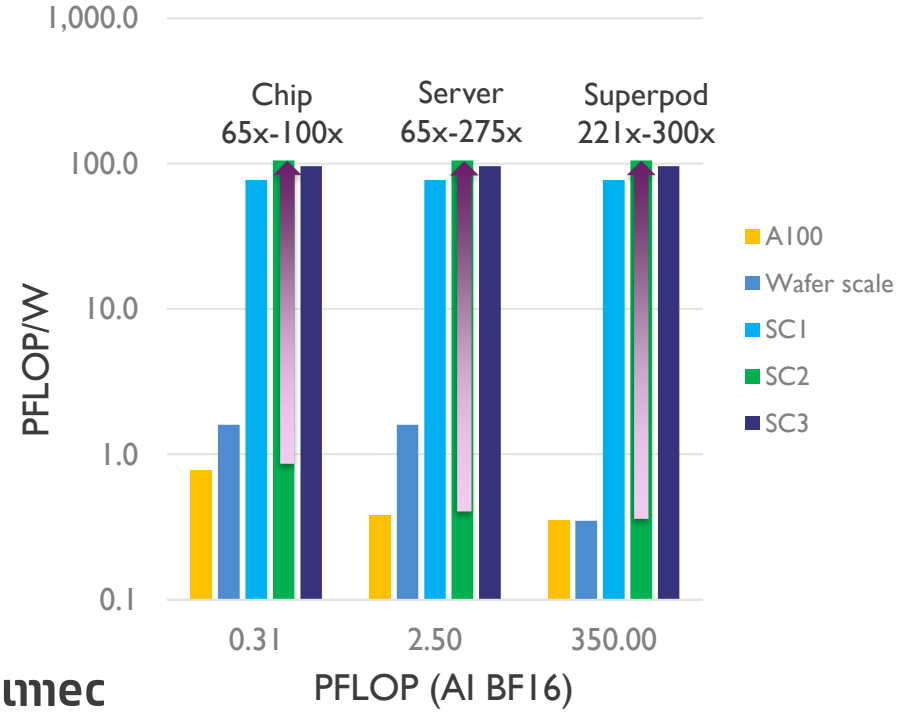


imec

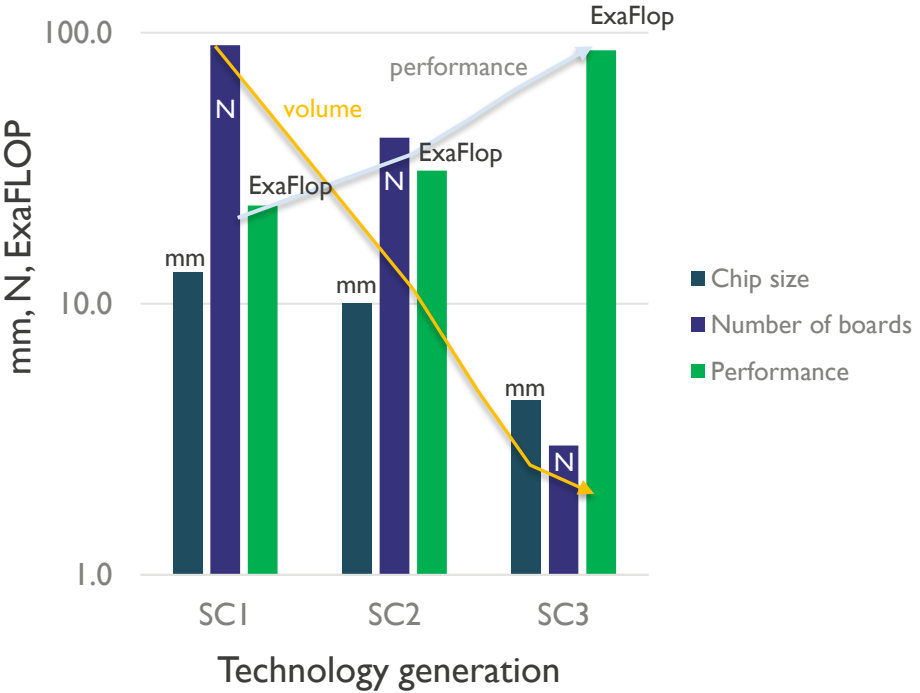
Efficiency and Computational Volume

Circuit energy efficiency is constant up to 100 ExaFlop

Energy efficiency vs. performance



Chip size, board count and performance



imec

Superconducting Digital Technology

Enabling sustainable hardware for deep learning and quantum computing

Greener

Smarter

More inclusive

100x energy efficiency

1000x compute density

Cheaper local systems


imec
embracing a better life