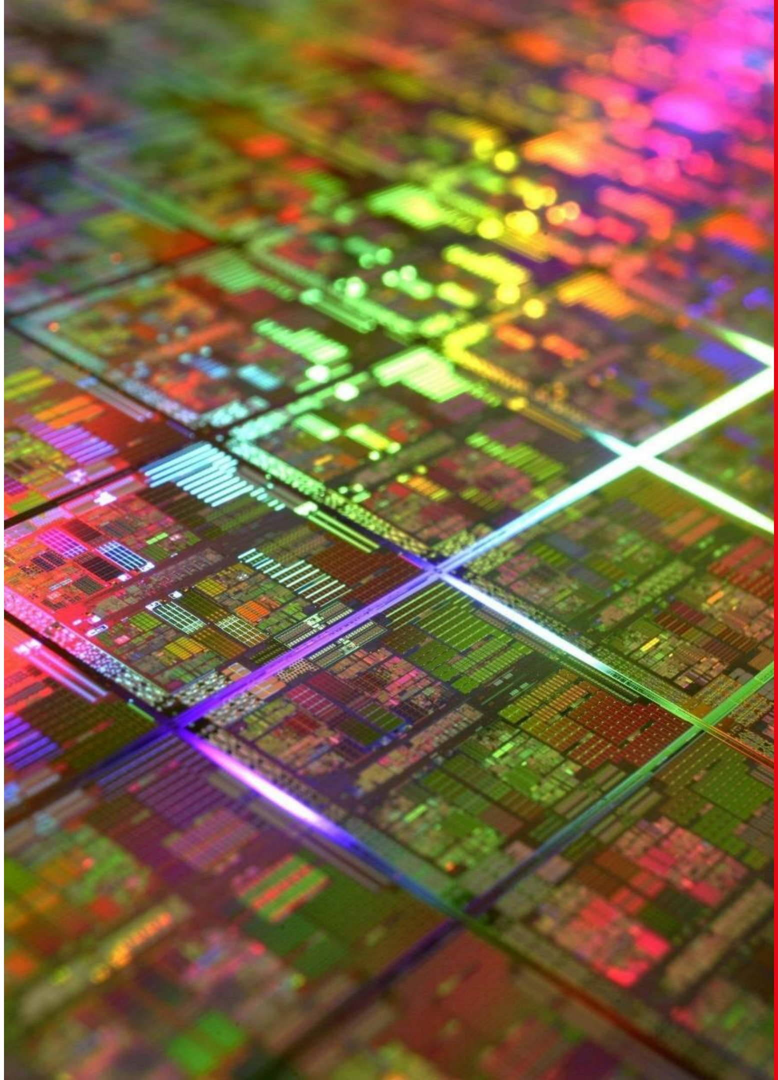


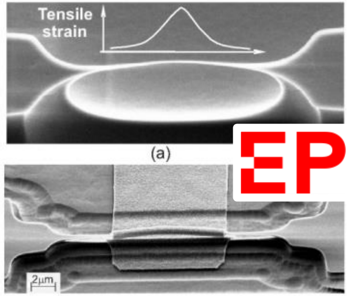
CMOS Emerging Architectures

Kirsten E. Moselund,

EPFL & PSI

A bit about my background…….

M.Sc in Engineering (Energy), DTU, Denmark



Ph.D in Microelectronics from EPFL, 2008
- Silicon CMOS, strain enhancement
Technology development, cleanroom

IBM Research Europe Zurich 2008- 2015, Post-doc, RSM
- Low-power electronics, III-V on Si integration for electronics.

IBM Research Europe – Zurich, 2016-2021,
Mgr. Materials Integration & Nanoscale Devices Group
~25 people, TEM & MOCVD facility, 5 PI’s
– Materials development, nanoscale thermometry, optoelectronic devices, Cryo electronics, Neuromorphic
- ERC StG on nanophotonics

My current position

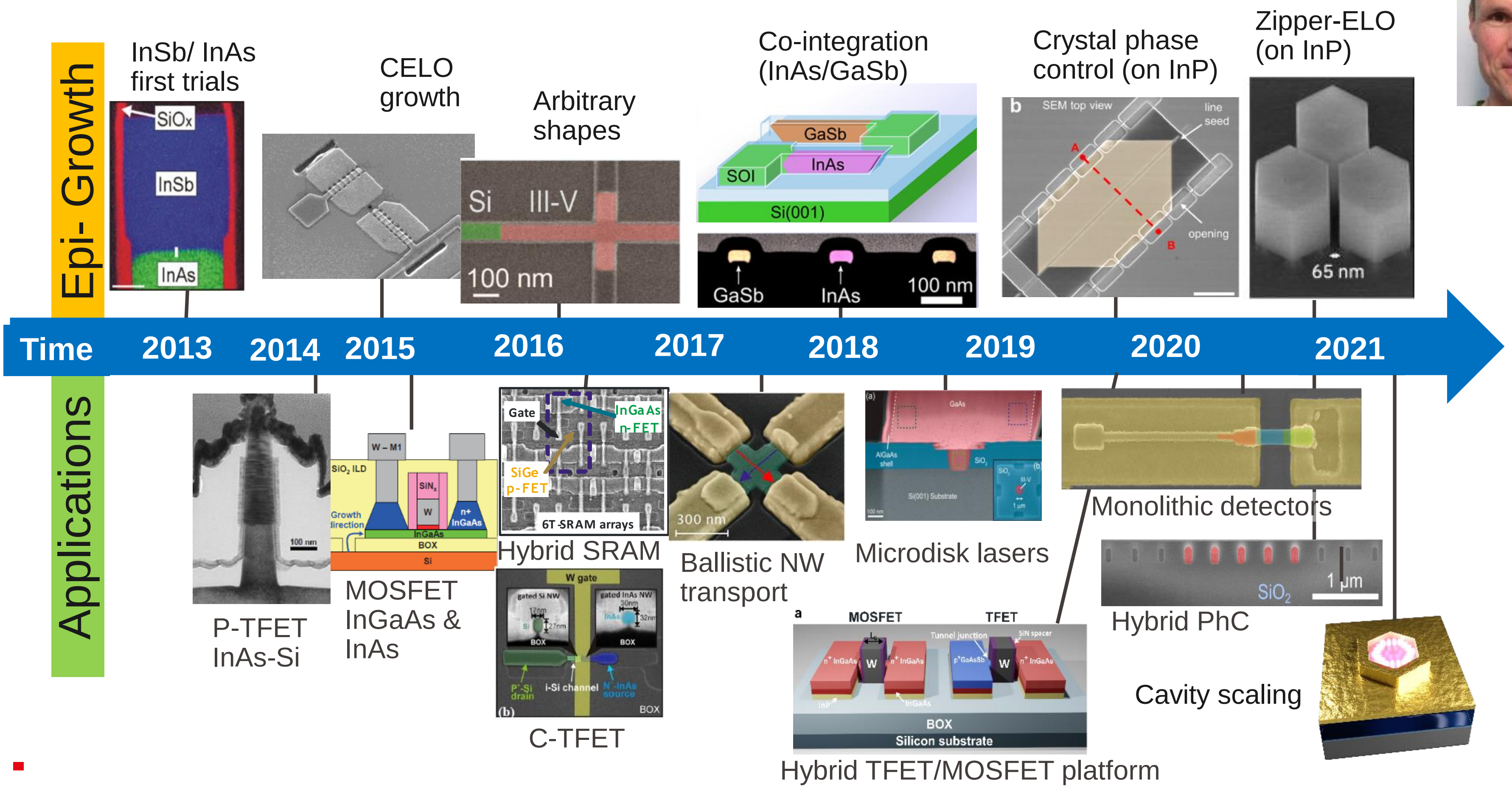
- **Professor at EPFL – 40%**
  – Integrated Nanoscale Photonics and Optoelectronics Laboratory (INPhO)
  – One day a week on average, usually Thursday or Friday
  – Teaching: MI 611 Nanoscale MOSFETs and beyond CMOS devices Fall, MI xx Nanophotonics from Spring 2023
  – Thesis advisor in EDMI
  – Part of QSE

- **Head of Laboratory of Nano and Quantum Technologies (LNQ) at PSI – 60%**

- **Temporary status of research activities**
  – Set-up and 4 students still at IBM, temporary solution until they finish their PhDs
  – Once Park Innovaare is finished will move set-up there
  – New activities will be principally located at PSI, students enrolled at EPFL

What I was working on recently

Heinz Schmid



PAUL SCHERRER INSTITUT

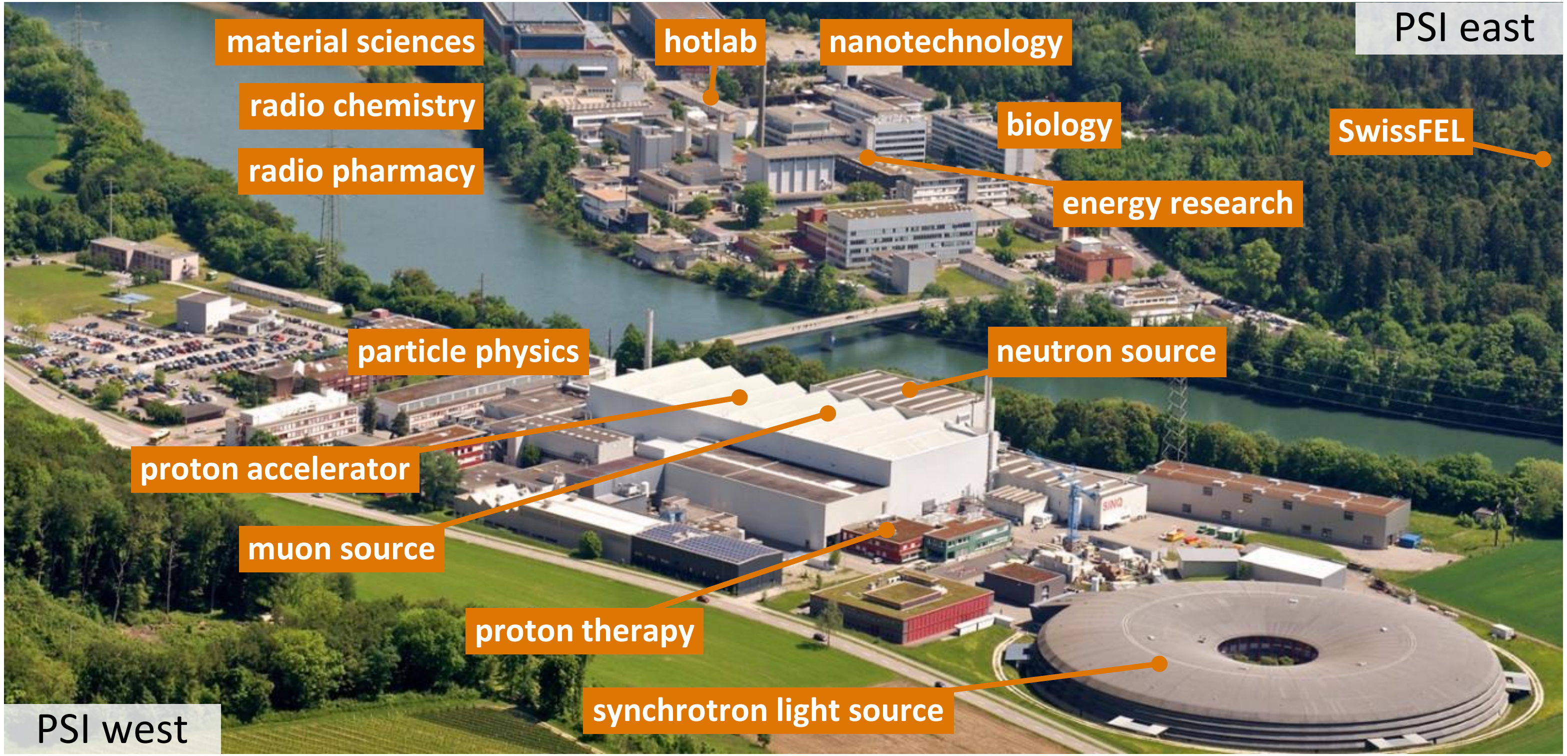
Overview

ETH Board

- **ETHZ** Swiss Federal Institute of Technology Zurich
- **EPFL** Swiss Federal Institute of Technology Lausanne
- **PSI** Paul Scherrer Institute
- **Empa** Swiss Federal Laboratories for Materials Testing
- **WSL** Swiss Federal Research Institute for Forestry, Snow and Landscape
- **Eawag** Swiss Federal Institute of Aquatic Science and Technology

← Basel

Germany ↑



Zürich →

Aarau/Bern ↓

Two Lectures in this course

Wednesday

10:45 – 12:15 CMOS emerging architectures

- Technology scaling
- Transistor architectures
- CMOS processing

Computation

Friday

14:00 – 15:30 Beyond CMOS: Integrated Photonics

- Interconnect bottleneck
- Optical communication
- Photodetectors and Emitters on silicon

Data transmission

Questions welcome anytime – please interrupt me!

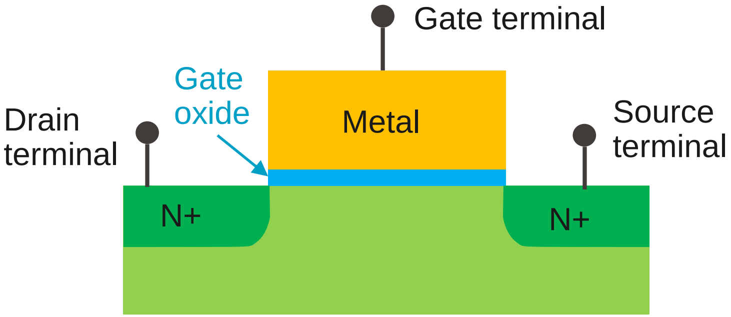
CMOS emerging architectures - Overview

- Technology scaling
  - Moore’s law – dead or alive
  - Technology nodes
- Transistor architectures evolution
  - Planar, FinFET, Nanosheet
- CMOS fabrication for advanced nodes

Basics: Metal-Oxide-Semiconductor Field Effect Transistor

Transistor = electronic switch

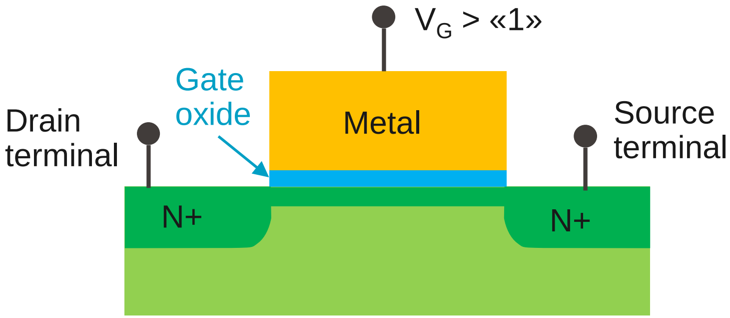
Transistor off – no path between source and drain



Only a very small leakage current flows.

Device off, logic “0”

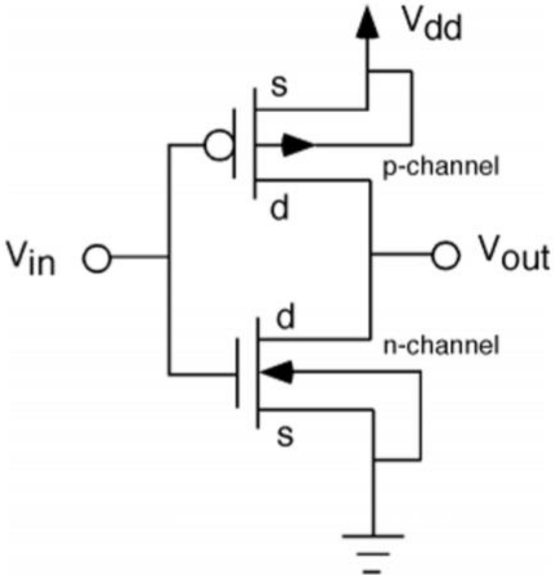
Transistor on – Conductive channel between source and drain



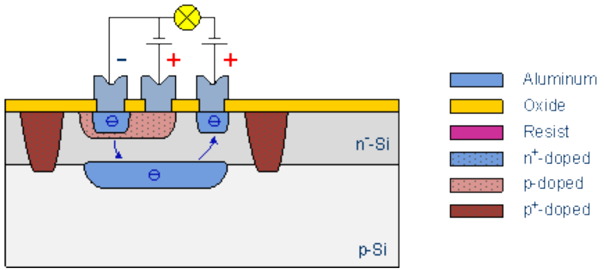
Current flows from S to D when there’s a difference in potential

Device on, logic “1”

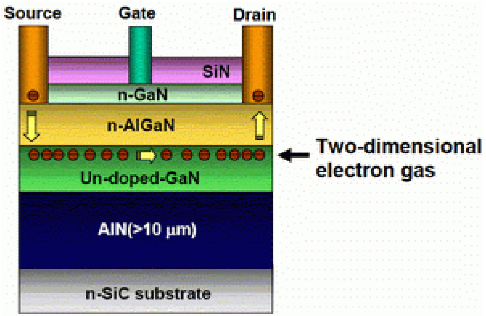
- Complementary MOSFET technology combines n-channel and p-channel device → lowest power consumption of logic designs
- Other transistors used for other applications



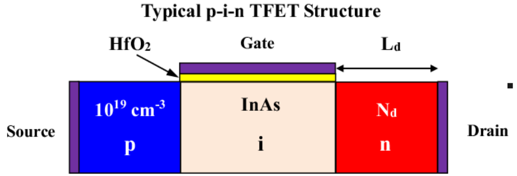
BJT – bipolar junction transistor



HEMT – high electron mobility transistor



TFET – tunnel FET

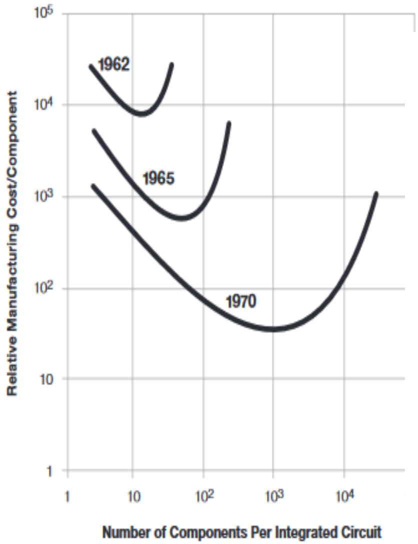


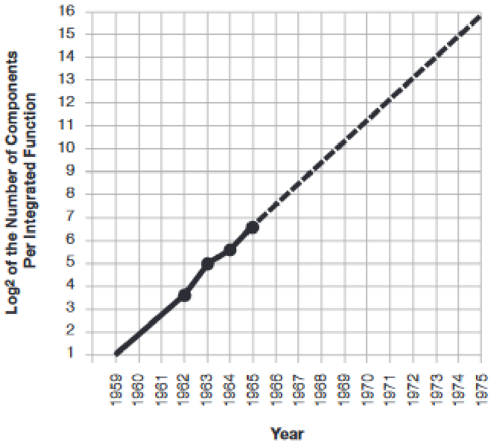
… and many more

This Lecture only deals with MOSFETs

Technology scaling

Moore's law – dead or alive





📖 G. moore, Electronics, Volume 38, Number 8, April 19, 1965

- Originally proposed in 1965 as a prediction of component cost

The complexity for minimum component costs has increased at a rate of roughly a factor of two per year (see graph on next page). Certainly over the short term this rate can be expected to continue, if not to increase. Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years. That means by 1975, the number of components per integrated circuit for minimum cost will be 65,000.

- Also interestingly predicted the dominance of silicon

Silicon is likely to remain the basic material, although others will be of use in specific applications. For example, gallium arsenide will be important in integrated microwave functions. But silicon will predominate at lower frequencies because of the technology which has already evolved around it and its oxide, and because it is an abundant and relatively inexpensive starting material.

Dennard scaling
The engineering aspect of Moore's law

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\epsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1

Table 2
Scaling Results for Interconnection Lines

Parameter	Scaling Factor
Line resistance, $R_L = \rho L/Wt$	κ
Normalized voltage drop IR_L/V	κ
Line response time R_LC	1
Line current density I/A	κ

- Proposed at IEDM 1972 (two years before the more cited paper)
- Lays out the rules governing transistor scaling
- Power density already identified as important metric
- Limitations
  - EOT (t_{ox}): already ~1nm, cannot scale atoms indefinitely
  - Operating voltage (V) scaling slowed, due to finite subthreshold swing (TFETs)

Effects of dimensional scaling

Classical scaling (Bulk)

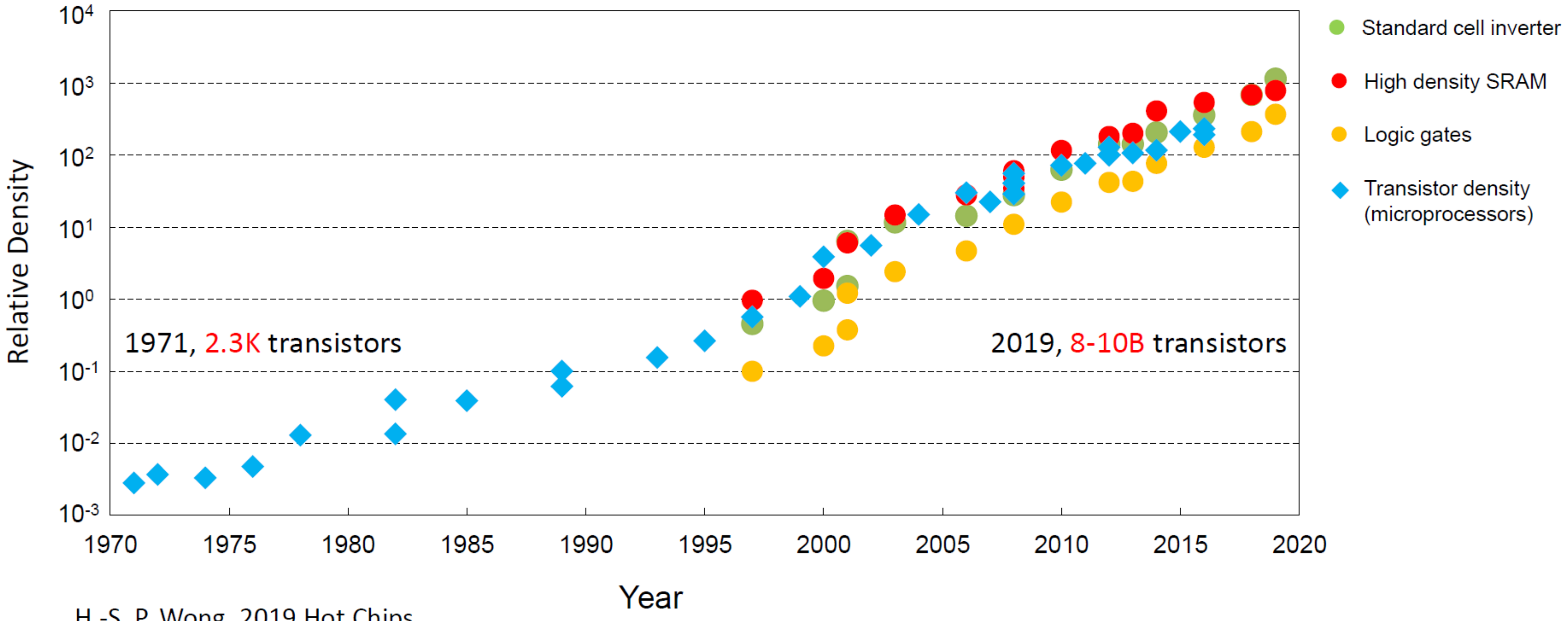
- Gate length getting shorter
- S/D contact area reduced
- BEOL – metal pitch is reduced
- → patterning (lithography) challenges

Performance scaling

- Focus: device speed and I_{on} at low power
- Concern of short-channel effects (SCE) and impact of capacitance (leakage ↑, speed ↓)
- Reduced contact area (resistance ↑, speed ↓)

Moore's Law – CMOS Density Improvement

Slides borrowed from Jin Cai, TSMC IEDM 2019

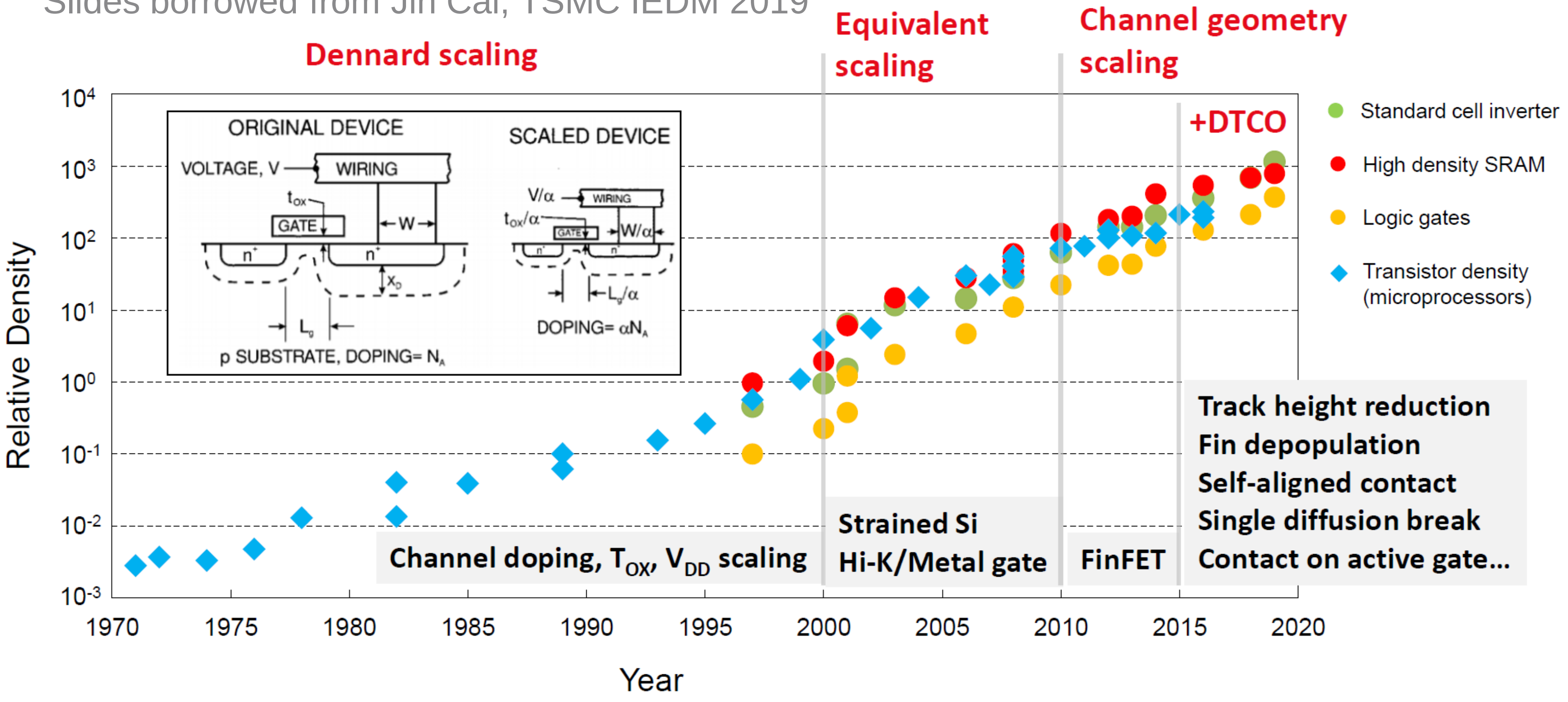


H.-S. P. Wong, 2019 Hot Chips

https://en.wikipedia.org/wiki/Transistor_count

Moore's Law – A History Of Innovations

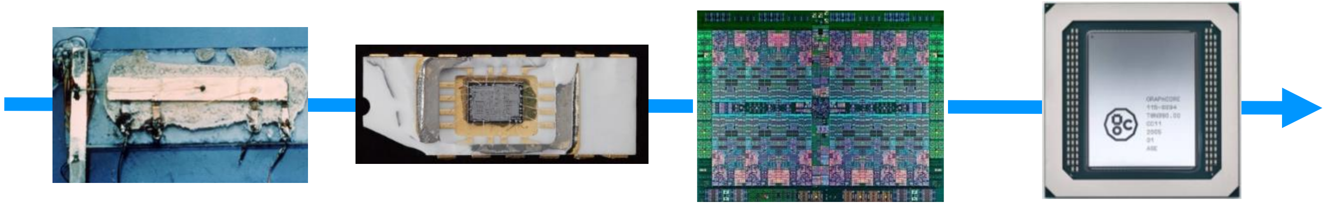
Slides borrowed from Jin Cai, TSMC IEDM 2019



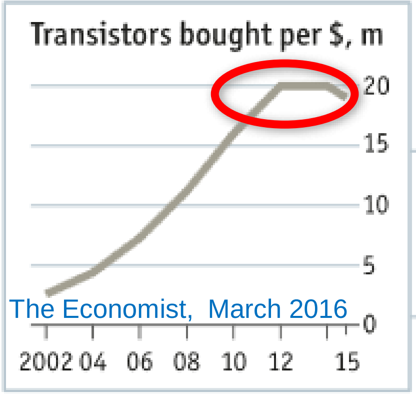
R. Dennard et.al., 1974 JSSC; D. Frank et.al., 2001 Proc. IEEE

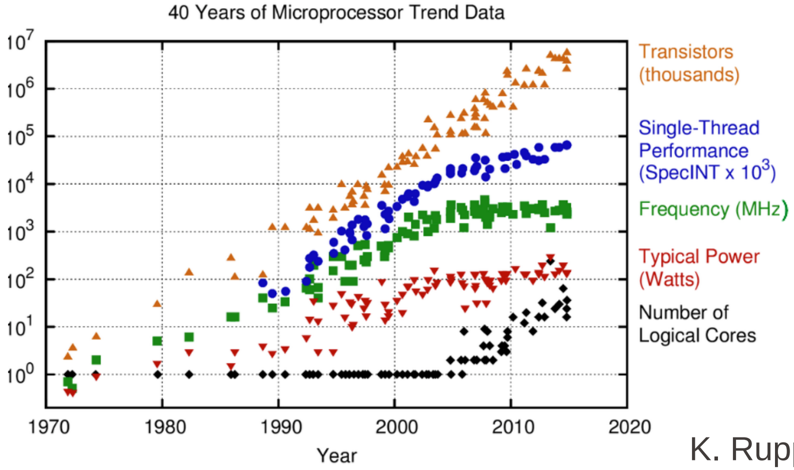
The Future of Computing – Next Generation Systems

1958	1971	2014	2020



First integrated circuit Size ~1cm² 2 Transistors	Moore's Law is under way Intel 4004 2,300 transistors	IBM P8 Processor ~ 650 mm² 22 nm, 16 cores > 4.2 Billion Transistors	Graphcore - GC200 IPU AI chip, 7 nm 1,472 cores 59.4 Billion Transistors

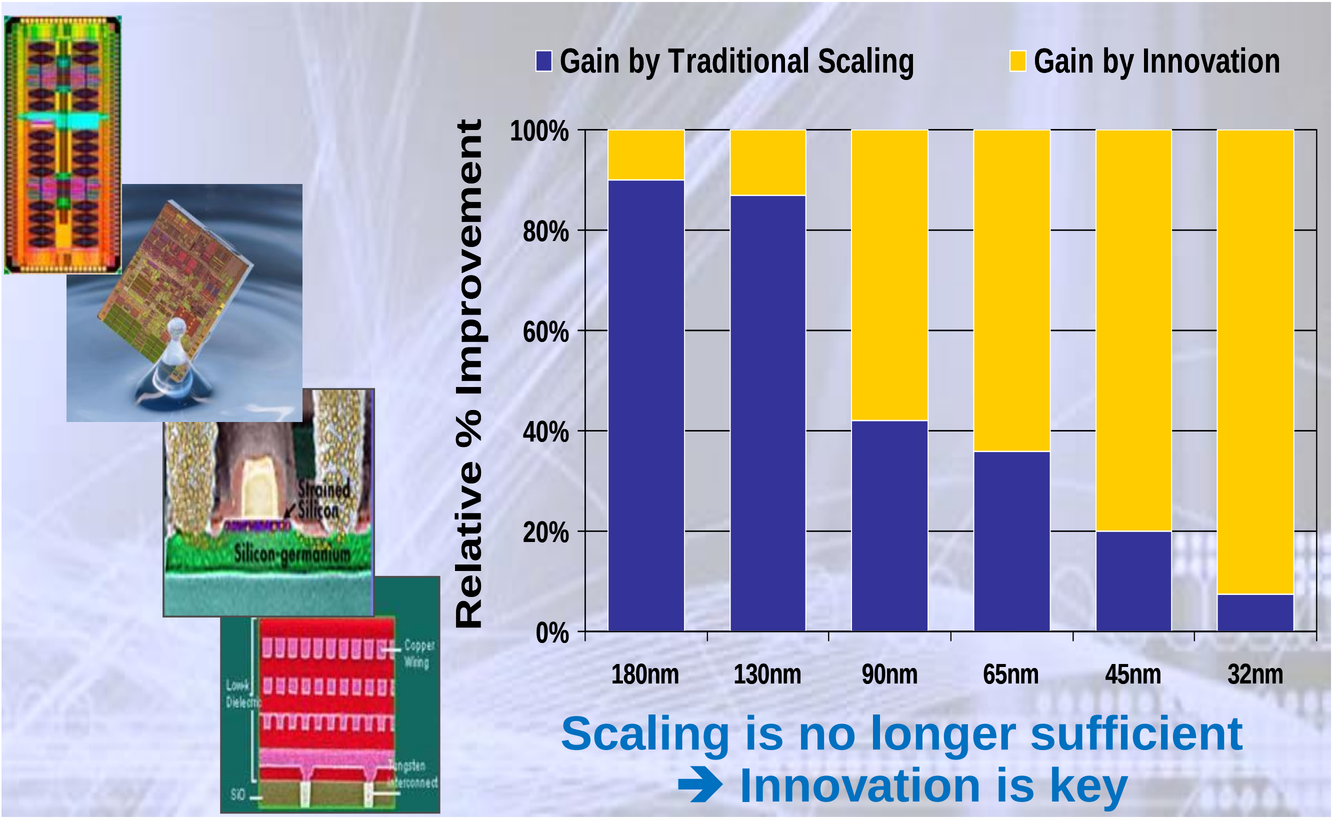




K. Rupp

- Complexity and density continues to increase
- Frequency and power need to flatten out
- Cost? – many flavors of devices, not all transistors are created equal

Innovation drives performance - not scaling



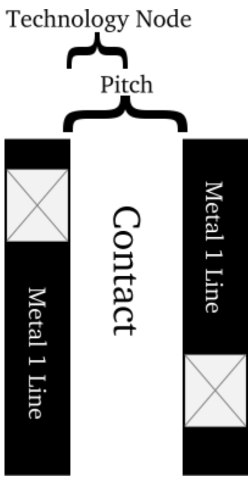
Technology node definition

Moore's Law: Density doubling → contacted poly pitch (CPP) and minimum metal pitch (MMP) need to scale by roughly 0.7x each node.

$$0.7x\ \text{CPP} \cdot 0.7x\ \text{MMP} \approx \tfrac{1}{2}\ \text{area}$$

1960s – 1990s

- Roughly L_G



1990s – 2000s

- More aggressive L_G scaling
- DRAM metal pitch drives technology node
- ITRS roadmap provides guidance

Since late 2000s

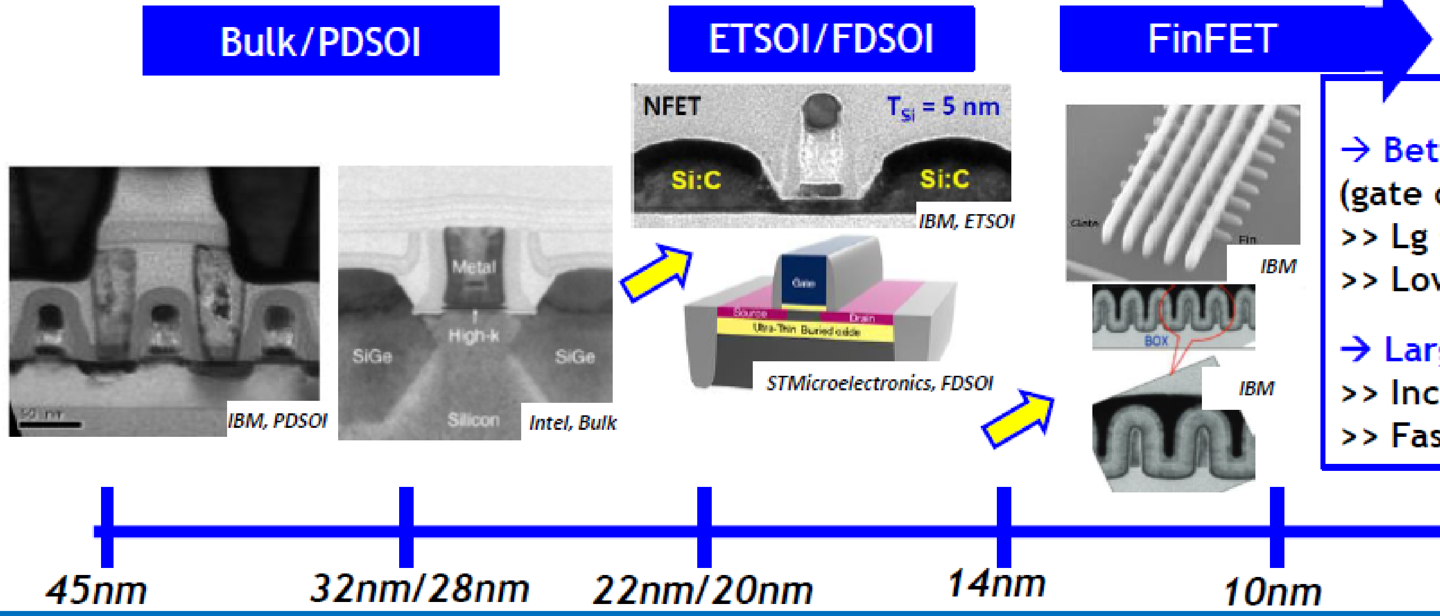
- L_G scaling no longer follows
- Emergence of new architectures →changes transistor scaling
- Recent nodes purely fictive representing a given generation of chips

Courtesy of: https://en.wikichip.org/wiki/technology_node

CMOS Device Performance by Architecture Engineering

📖 N. Loubet , IBM, VLSI 2020

- Historical move from a classical Planar device to transistors architectures with extremely thin channel: FDSOI Transistors covering CMOS Logic in the 32nm/22nm device nodes
  Significant breakthrough with the insertion of 1st FinFET technology at 14nm



FinFET vs. Bulk:

→ Better SCE
(gate over 3 sides of Fin)
>> Lg scaling
>> Lower leakage

→ Larger Weff/active footprint
>> Increase in Performance ↑
>> Faster at lower Power

EPFL

Comparing Intel 10nm with Samsung & TSMC 7nm nodes

Many specs. are similar, possibly different target groups - denser SRAM pitch

		Intel		TSMC		Samsung	
	Process Name	P1274 (CPU) / P1275 (SoC)		7FF, 7FF+, 7HPC		7LPE	
	1st Production	2018		Q1, 2018		2019	
Lithography	Lithography	193 nm		193 nm		EUV	
	Immersion	Yes		Yes			
	Exposure	SAQP		SAQP		SE	
Wafer	Type	Bulk		Bulk		Bulk	
	Size	300 mm		300 mm		300 mm	
Transistor	Type	FinFET		FinFET		FinFET	
	Voltage	0.70 V		0.70 V			
		Value	14 nm Δ	Value	10 nm Δ	Value	10 nm Δ
Fin	Pitch	34 nm	0.81x				
	Width	7 nm	0.88x	6 nm	1.00x		
	Height	53 nm	1.26x	52 nm	1.24x		
	Gate Length (L_g)						
	Contacted Gate Pitch (CPP)	54 nm	0.77x	55 nm	0.84x	54 nm	0.79x
	Minimum Metal Pitch (MMP)	36 nm	0.69x	40 nm	0.95x	36 nm	0.7x
SRAM bitcell	High-Perf (HP)	0.0441 µm²	0.62x				
	High-Density (HD)	0.0312 µm²	0.62x	0.027 µm²	0.64x	0.0260 µm²	0.65x

Courtesy of: https://www.extremetech.com/computing/296154-how-are-process-nodes-defined

Technology as a marketing booster

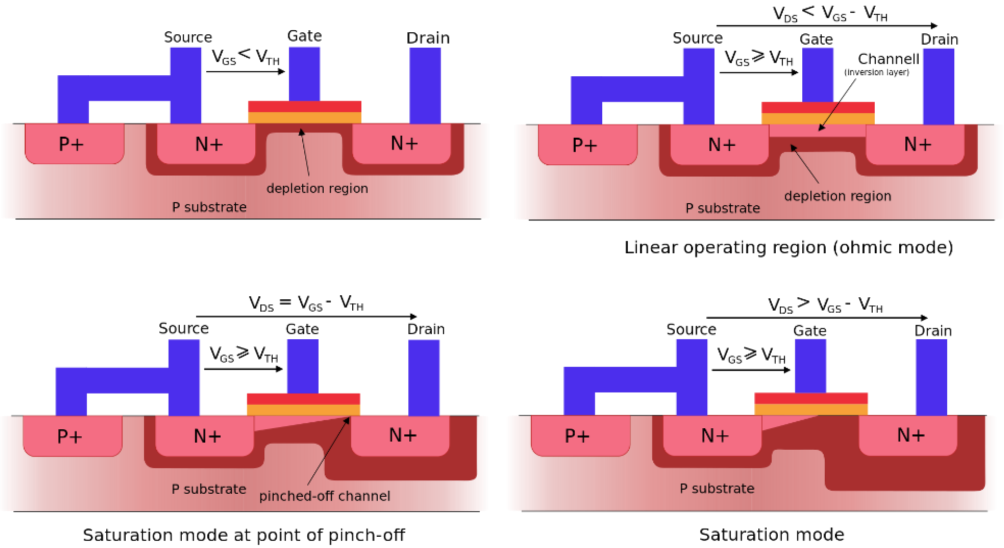
Transition from an actual technology definition as financial driver for the semiconductor industry, to a marketing booster vaguely implying enhanced performance

Partial Summary: Scaling

- Moore’s law in terms of device density and functionality continues.
- Frequency scaling does not – because of power concerns, and multi-core architectures
- The technology node has lost its meaning in terns of a specific device dimension, but vaguely defines the next step in technology processes
- Hard to compare different technology flavors

Transistor architectures

Operating principle of the MOSFET



📖 O. Deleage and P. Scott, Wikipedia MOSFET

- Killer advantage is that almost no current is required to modify the channel conduction – unlike a BJT.
- Functionality: Modification of the conduction of the channel by the application of a gate bias
- The better the electrostatic coupling of the gate to the channel the more efficient the operation
- Readily lends itself to scaling and VLSI fabrication

Important MOSFET figures of merit

- **Drive Current:** Saturation regime $I_{D,sat} = \frac{\mu C_{ox}}{2} \cdot \frac{W}{L} \cdot (V_{GS} - V_T)^2$
- **Mobility**, $\mu = \frac{q\tau}{m_{eff}}$ → the speed with which free carriers move in an E-field is inversely proportional to the effective mass
- **Gate oxide capacitance:** $C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{k_{ox}\epsilon_0}{t_{ox}}$, This is the capacitance which controls the amount of charge in the channel → wants to increase it → high-k dielectrics. Similar scaling applies to parasitic capacitances, but I those cases we want to reduce (low-k dielectrics or air)
- **Subthreshold swing:** steepness of the on-off transient of the transistor. Ideal case *60mV/dec@ 300K*. The smaller the slope the less voltage is needed to switch the device on/off.
- **Power consumption:** $P_{total} \approx CV_{dd}^2 f(V) + I_{Leak}V_{dd}$

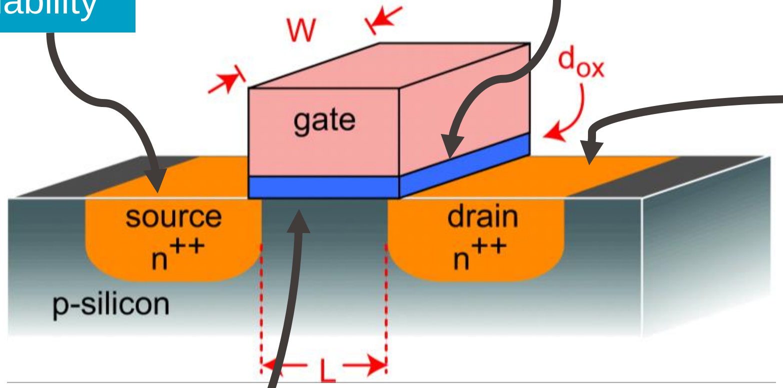
kirsten.moselund@epfl.ch

Scaling Limitations of the MOSFET

d_{ox} si reduced
→ Increasing gate leakage

Finite # of dopants
→ device variability

Surface area of contact region is reduced → R_c ↑

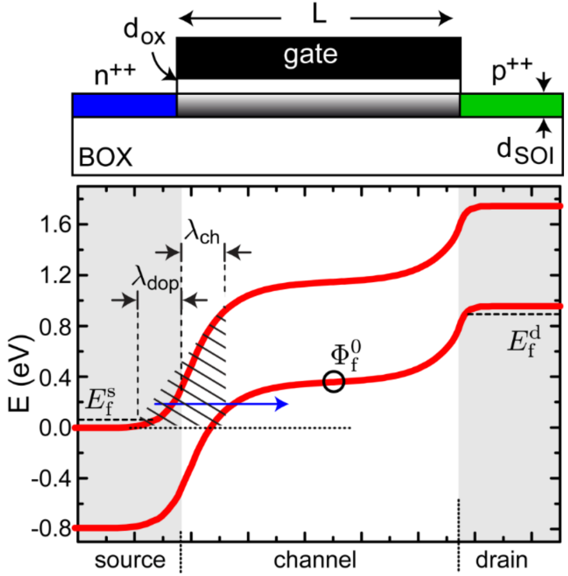


L is reduced → fraction of Q controlled by gate decreases

V_T shift, DIBL, and increased inverse subthreshold slope

The Screening Length, λ

J. P. Colinge SSE vol 48, 2004



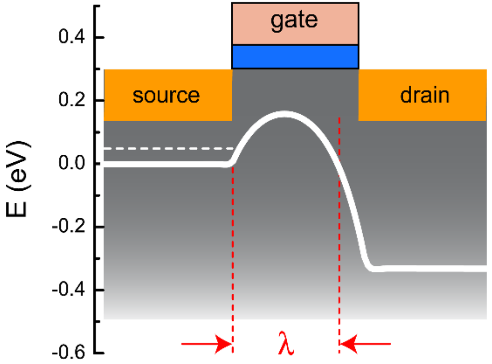
Encroachment of the electrical field from the drain, reduces the effective charge controlled by the gate, and reducing V_T.

The *screening length* or the *natural length*, λ, represents the penetration of the electric field lines from the drain into the channel.

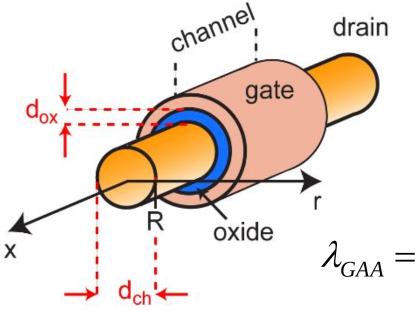
- To avoid being limited by SCE L_G should be greater than 5-10x λ.
- Depends on geometry, ε_{ox}, t_{ox} and t_{NW}.

EPFL

Comparison of Planar vs. Nanowire Architecture



$$\lambda = \sqrt{\frac{\varepsilon_{Si}}{\varepsilon_{ox}} t_{ox} t_{Si}}$$



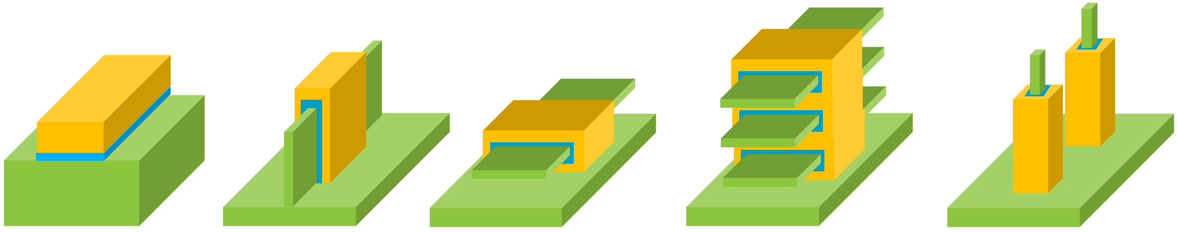
$$\lambda_{GAA} = \sqrt{\frac{2\varepsilon_{NW} t_{NW}^2 \ln\left(1+\frac{2t_{ox}}{t_{NW}}\right) + \varepsilon_{ox} t_{NW}^2}{16\varepsilon_{ox}}}$$

- **Planar gate:** limited electrostatic control of the channel
- **<u>Example:</u>** 8nm SOI, 1 nm SiO_2: λ ≈ 5 nm → **L_G > 25 nm**

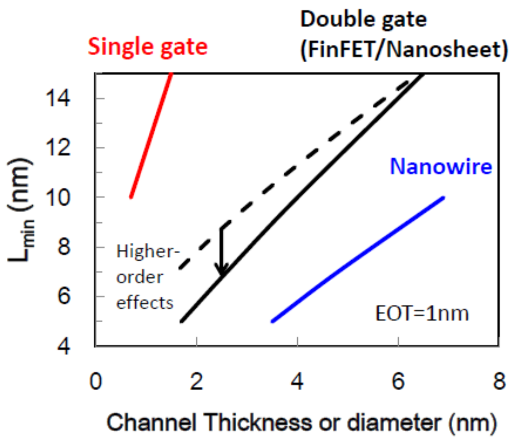
- **Surround gate (Nanowire):** ultimate electrostatic control of channel
- **<u>Example:</u>** 8 nm SiNW, 1 nm SiO_2: λ ≈ 3.1 nm → **L_G > 15 nm**

NW device geometry yields improved scaling and better subthreshold swing

Device options for gate length scaling



Planar | FinFET | Nanosheet | Stacked nanosheet | Vertical nanowire



Solid lines: TSMC Simulation
Dash line: Scale length theory, Frank, Taur, Wong, 1998 EDL

Jin Cai, TSMC IEDM 2019

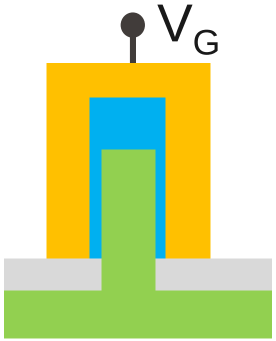
Improving electrostatics → L_G scaling

Improved Lay-out efficiency

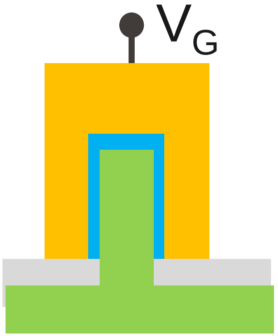
Improved Lay-out efficiency

Exploiting the vertical direction

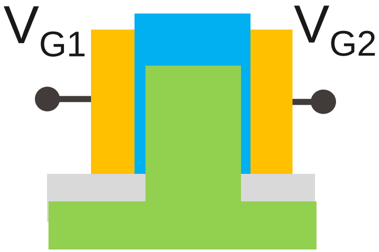
FinFET architecture



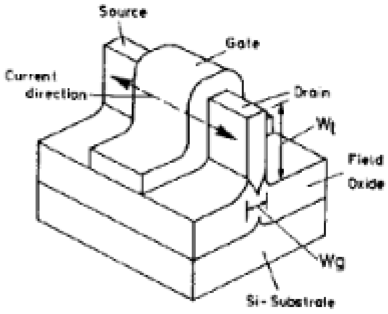
Double gate



Tri-gate

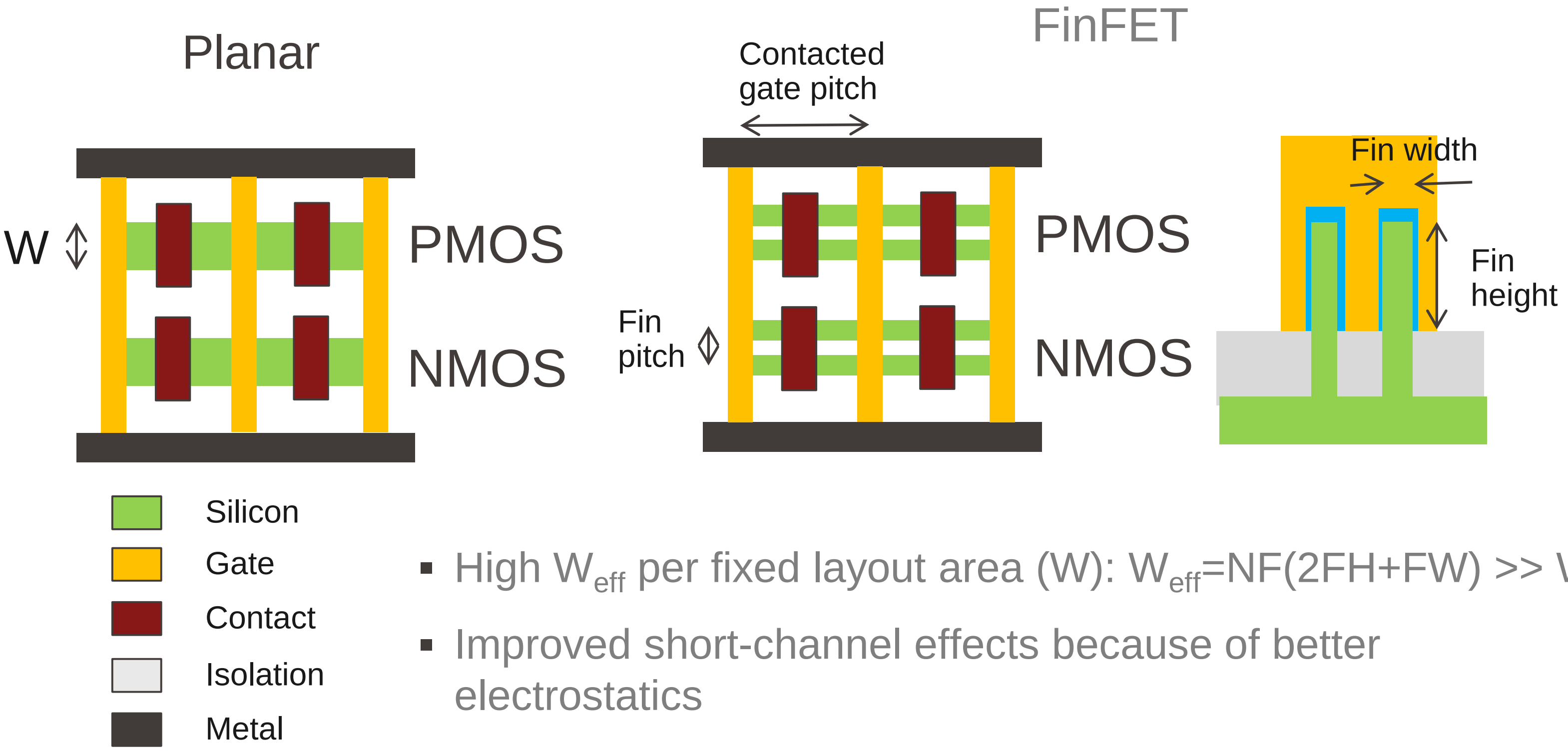


Independent double gate



- 1989 proposed by Hitachi at IEDM (Hisamoto et al.) - depleted lean-channel transistor (DELTA). A decade of intense university research follows.
- 2002: First 25 nm Ω-FinFET operating on just 0.7 volt demonstrated by TSMC
- Commercial production of nanoelectronic FinFET semiconductor memory started in 2010s
- 2012 Intel began releasing Tri-gate CPU technology
- Commercially produced CPU chips at 22 nm and below have utilised FinFET gate designs

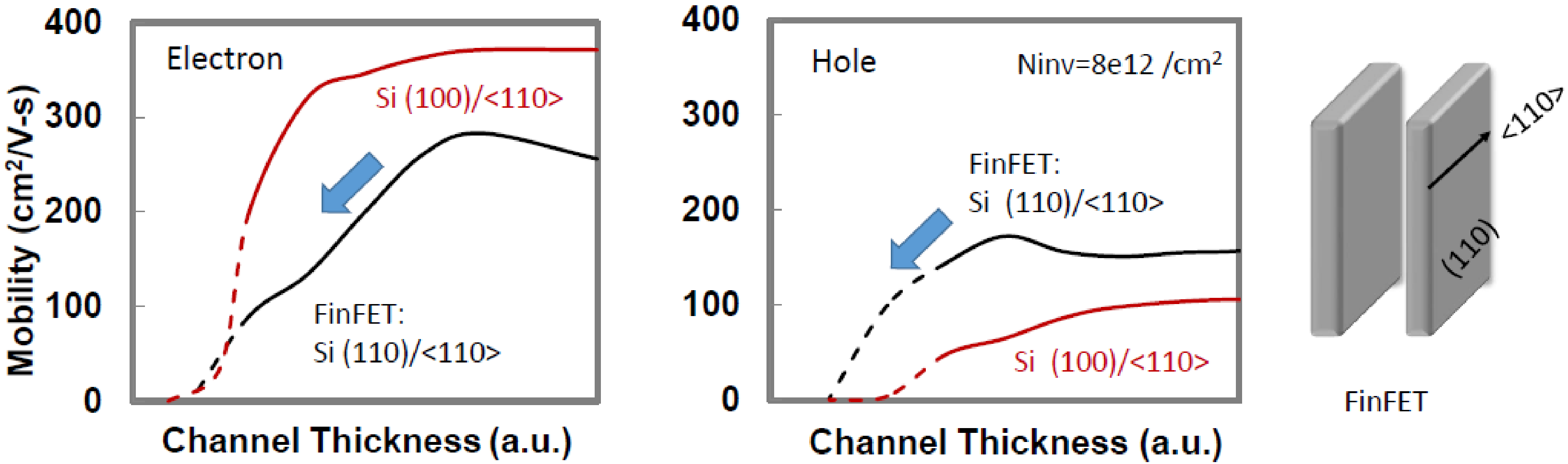
finFET → mproved lay-out efficiency



- High W_{eff} per fixed layout area (W): W_{eff}=NF(2FH+FW) >> W
- Improved short-channel effects because of better electrostatics

Performance optimization

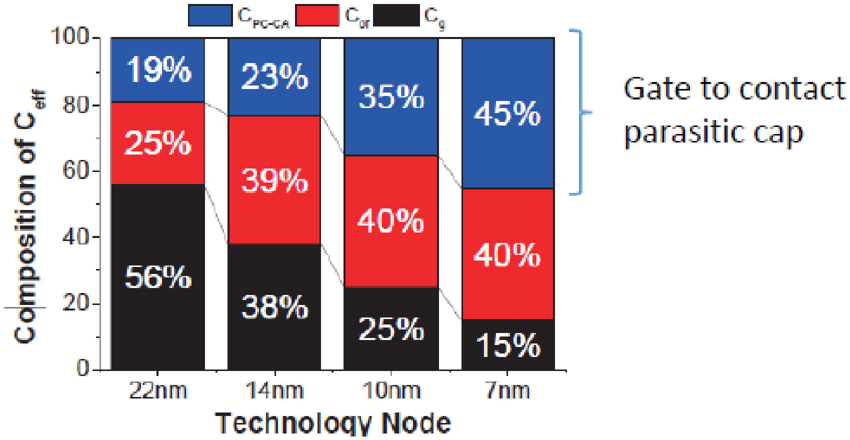
- Short-channel effects improve with diminishing fin width
- But, mobility degrades because of sidewall roughness scattering.
- Quantization also plays a role for thin fin widths
- Complex optimization as a function of wafer and fin orientation.



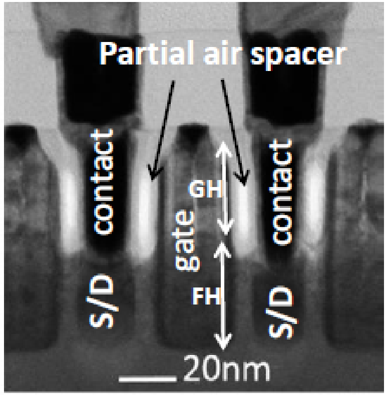
- 📖 Jin Cai, TSMC IEDM 2019

Parasitic capacitance reduction

- Gate to S/D contact cap can be reduced by low-K or air spacer
- Benefit of partial air spacer is limited with gate height reduction
  - Full air spacer is desirable, need high-selectivity nitride spacer removal



T. Yamashita et al., 2015 VLSI (IBM)



K. Cheng et al., 2016 IEDM (IBM)

GH=gate height; FH=fin height

Jin Cai, TSMC IEDM 2019

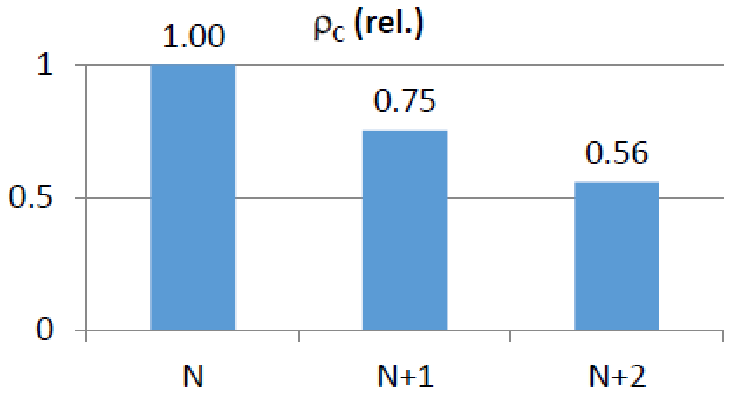
Contact Resistance Reduction

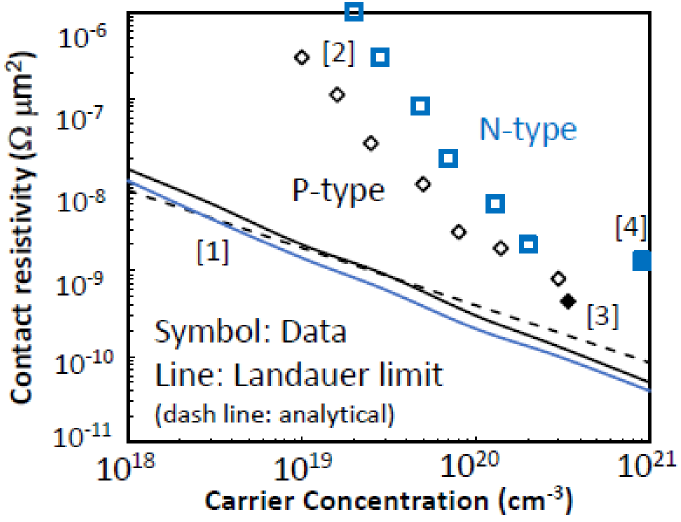
- Strategy: (1) reduce ρ_C (2) Increase contact area
- Fin pitch and gate pitch reduction reduces contact area
  - Contact resistivity (ρ_C) reduction necessary to keep Rc low
  - *Plenty of room to the theoretical lower limit*

$$R_C = 2(\rho_C/A_{con})W_{eff}$$

(short contact limit)

$$R_c = \frac{2\rho_C}{L_{CON}FP}(2FH + W_{FIN}) = 50\ \Omega\ \mu m$$

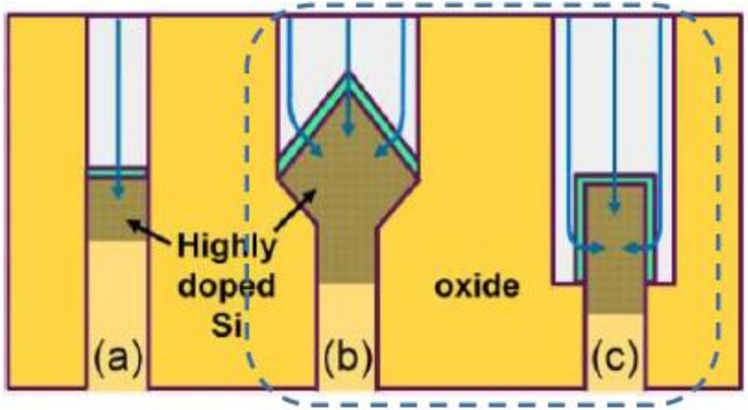




[1] J. Maassen et al., 2013 APL (Purdue): theoretical limit
[2] N. Stavitski et al., 2008 TED (NXP): Si data
[3] Y. Wu et al. (NUS), 2018 VLSI (NUS): GeSn data
[4] H. Yu et al., 2016 TED (IMEC): Si data

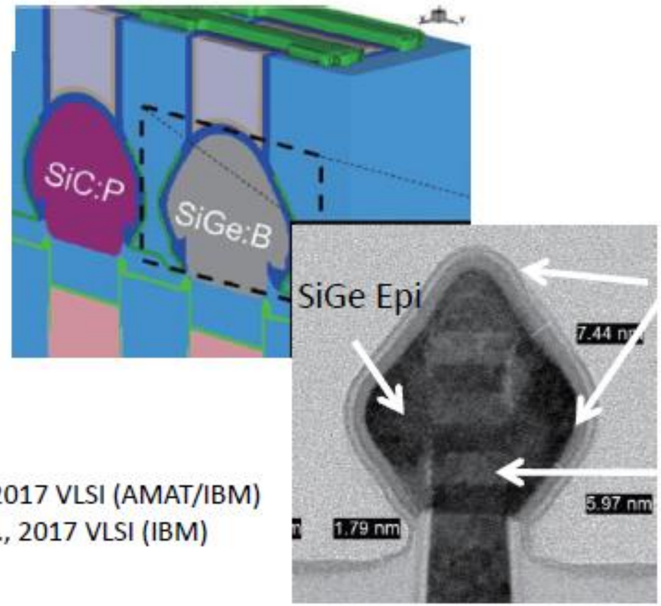
Maximize Contact Area

- Nominal contact area reduces with CGP and fin pitch scaling
- Wrap-around contact (WAC) can maximize S/D contact area



$$R_{con} \propto \frac{\rho_C}{Acon}$$

H. Yu et al., 2016 TED (IMEC)

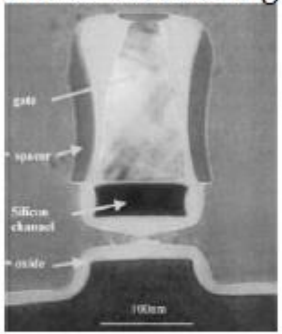


N. Breil et al., 2017 VLSI (AMAT/IBM)
N. Loubet et al., 2017 VLSI (IBM)

Gate-all-around (GAA) or Nanosheets

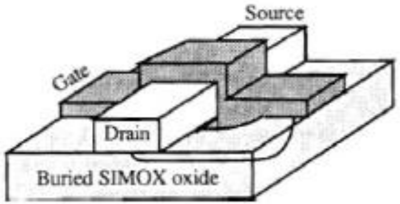
- The wrap-around gate architecture provides the optimum electrostatic control
- 1990: Proposed J.P. Colinge
- 2000's Silicon-on-nothing developed by ST Microelectronics
- Stacking of nanosheets allows for improved lay-out efficiency – more current than a single NW
- Improved SCE compared to FinFET only for thin sheets

silicon-on-nothing



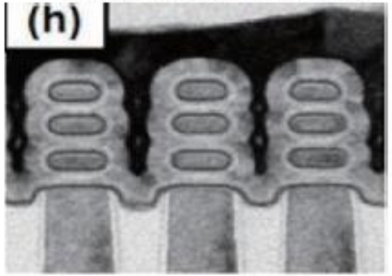
M. Jurczak et al., 2000 TED

'GAA'

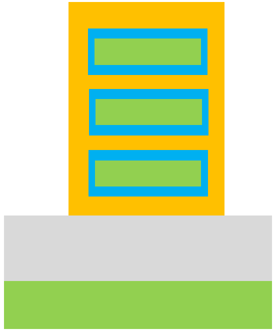


J.P. Colinge et al., 1990 IEDM

Stacked nanosheet



N. Loubet et al., 2017 VLSI



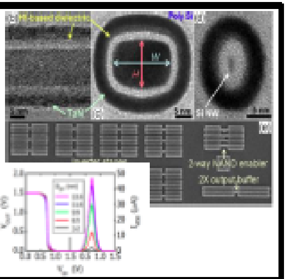
N. Loubet , IBM, VLSI 2020

Gate-All-Around at IBM Research

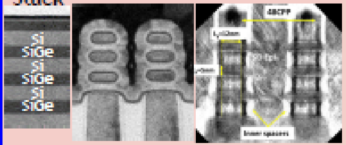
VLSI 2010

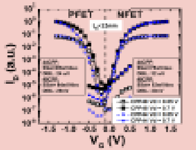
1st Demo of 300 mm Gate first CMOS Si NW.

1st Demo of Si NW Ring Oscillators.



VLSI 2017

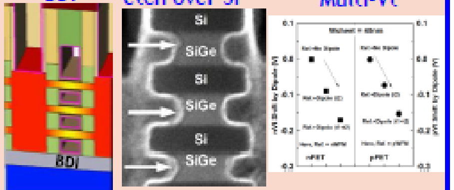




3 stack devices
Lg=12nm
44/48nm CPP

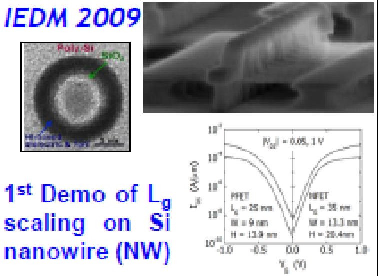
1st Demo of Stacked GAA NS at 7nm ground rules

IEDM 2019



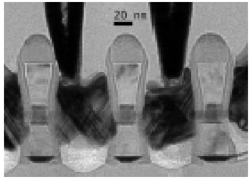
2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020

IEDM 2009



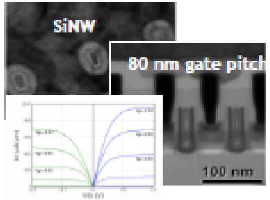
1st Demo of L_g scaling on Si nanowire (NW)

IEDM 2013



1st Demo of Si NW FET at 60nm gate pitch.

VLSI 2015

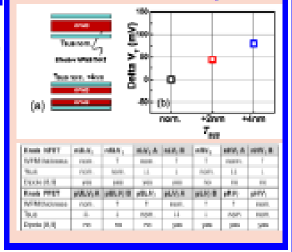


1st Demo of 300 mm RMG CMOS Si NW at 14 nm ground rules.

Evolution from <u>single wire</u> to <u>stacked sheets</u>

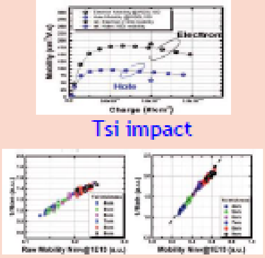
IEDM 2017

Nanosheet
Multi-Vt -Tsus impact



IEDM 2018

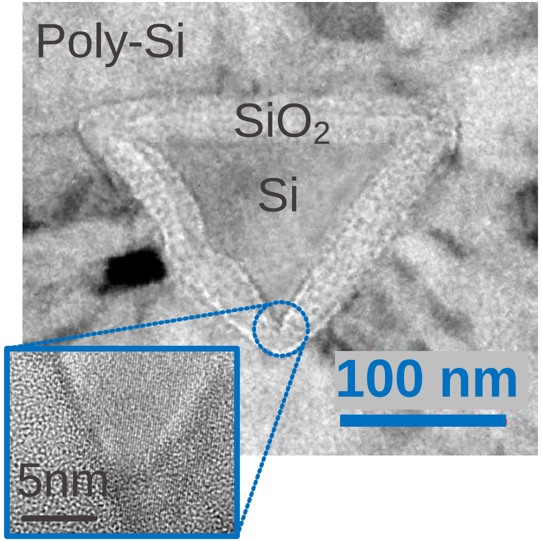
Nanosheet
Mobility

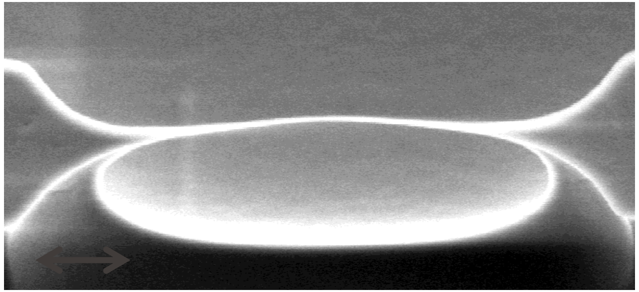


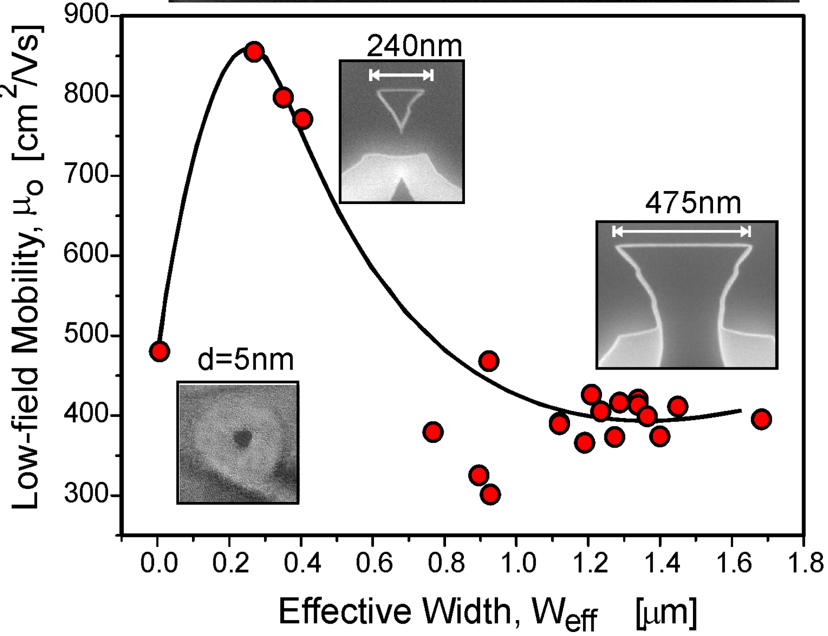
Tsi impact

- 15 years of R&D on horizontal Gate-All-Around (h-GAA) Architecture.
- Device architecture evolved from a <u>single Nanowire</u> to <u>stacked Nanosheet</u> for technology competitiveness over FinFET: larger Weff/footprint and reduced capacitance

My thesis work at EPFL – Strained GAA FETs (2008)





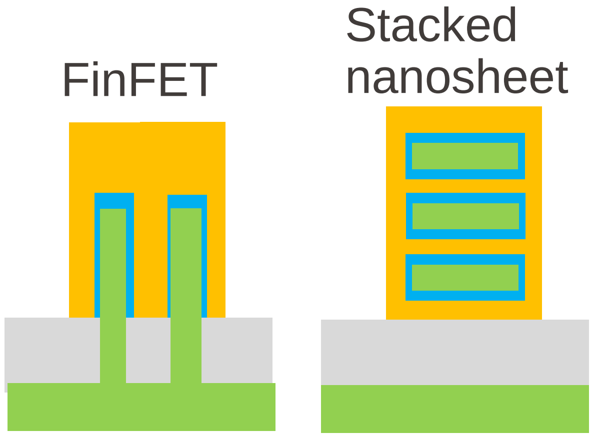


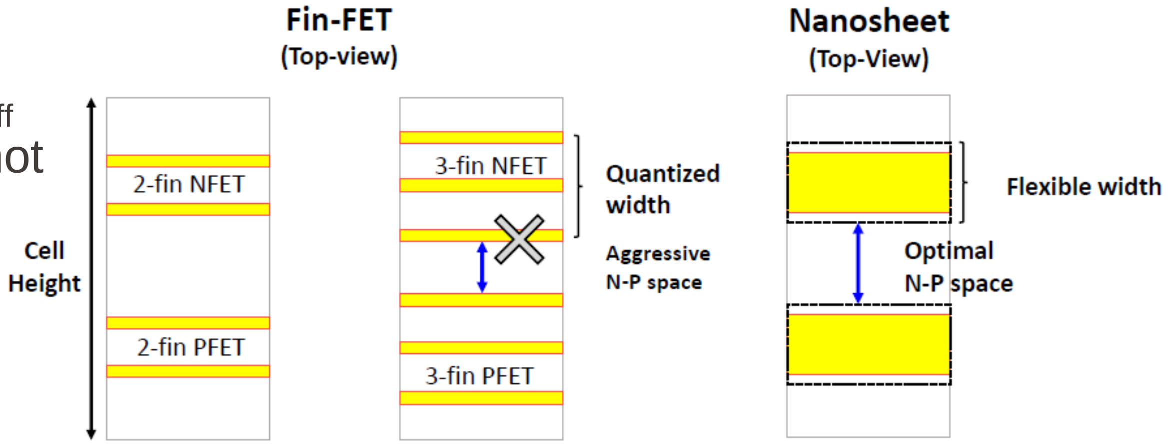
- Dimensional scaling by self-limited oxidation
- Oxidation causes strain in NW → boosting mobility
- Triangular GAA or W-gate architectures

Comparison: FinFET vs. Nanosheets

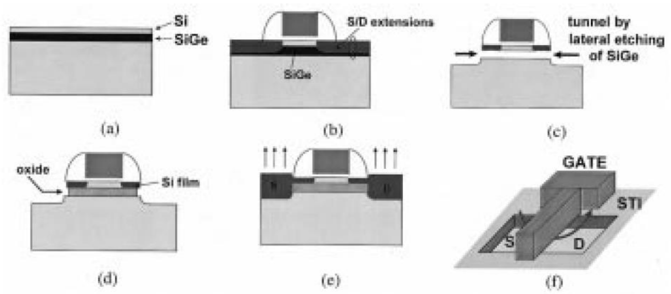
- Stacking of nanosheets allows for improved lay-out efficiency
- Improved SCE compared to FinFET only for thin sheets

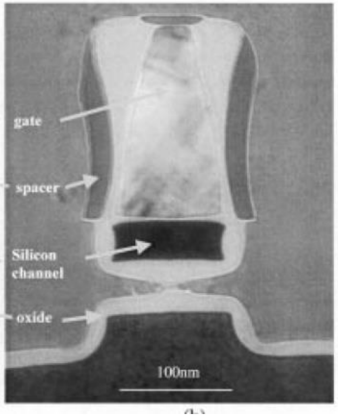
- Continuous width adds flexibility in design
- Width scaling limited by I_{eff}/C_{eff} performance trade-off → cannot make sheets arbitrarily wide

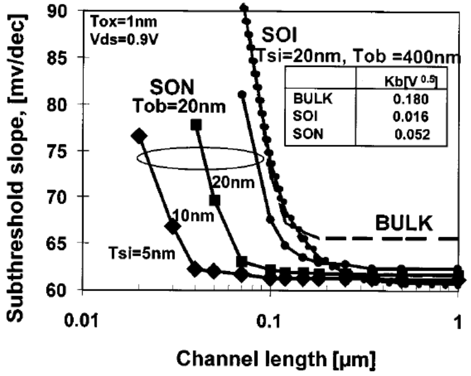




Silicon-on-Nothing: ST Microelectronics – 2000s







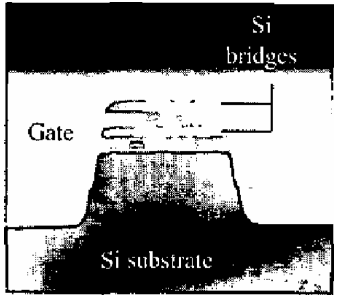
- The use of Si/SiGe superlattices is the key to stacked nanosheet processes.
- A controlled process which lends itself to mass production → possibility for multiple sheet stacking is evident.
- First device only top-gate
- Better performance than bulk and SOI
- Substantially improves the subthreshold swing
- Non-lattice matched stack

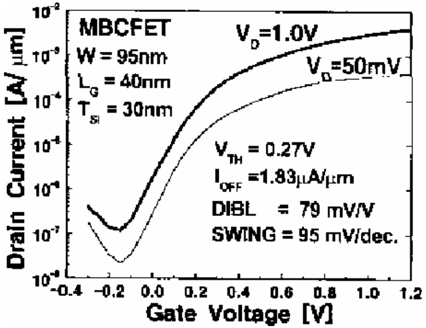
- 📖 M. Jurczak et al. TED, 2000

EPFL

Evolution towards stacked nanosheet

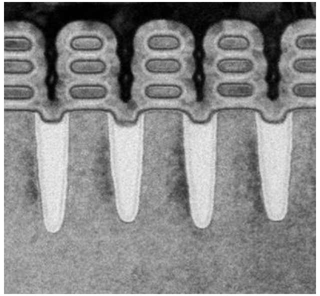
- 2004: Samsung: multi-bridge-channel MOSFET (MBCFET)
- Two stacked nanosheets
- Comparing single and double sheet performance

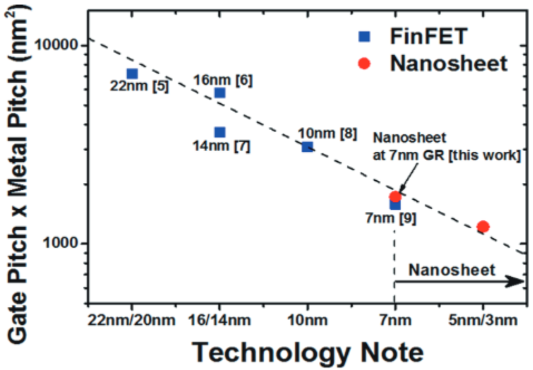




S.Y. Lee et al. VLSI Symp. 2004

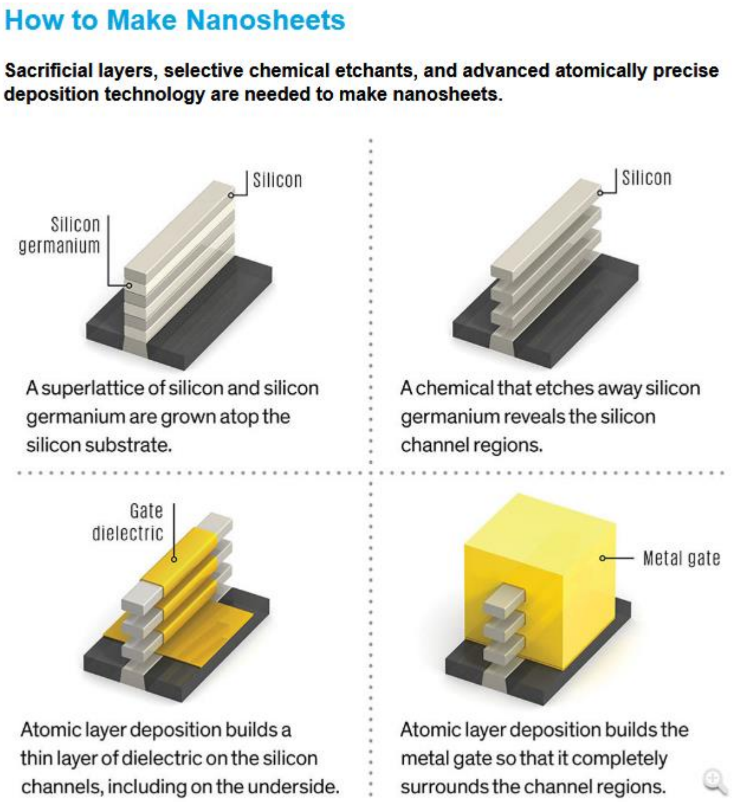
- 2017: IBM demonstrates NS for 7nm and beyond ground rule
- Stacks of three with sizes ranging from 8 to 50 nm across.





N. Loubet et al. VLSI Symp. 2017

Basic nanosheet process

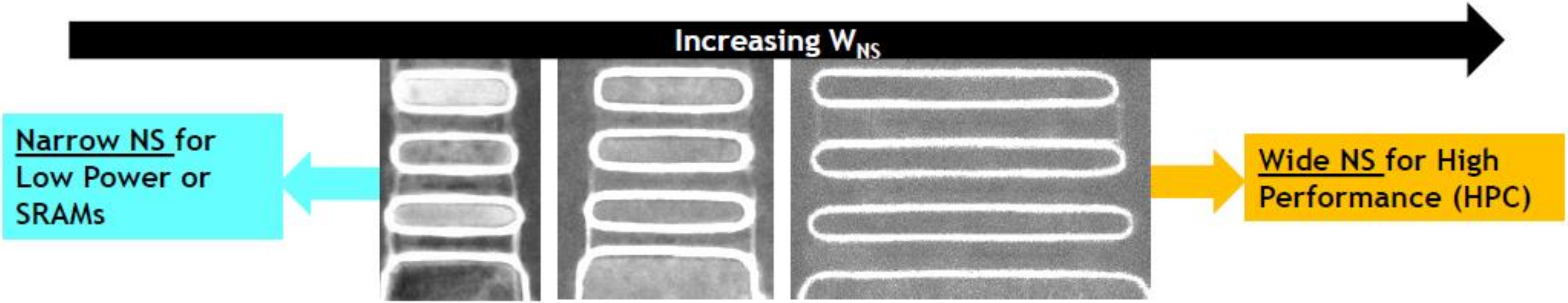


P. Ye et al. IEEE Spectrum, 2019

- Superlattice – Si/SiGe (mechanical stresses and capacitances limits number of layers)
- Selective chemical etching of SiGe → leaves Si nanosheets suspended as bridges between source and drain
- Atomic-Layer-deposition (ALD) → fills the gap with high-k and gate metal.

SiGe etch for channel release

- Key feature: variable width nanosheet on same wafer

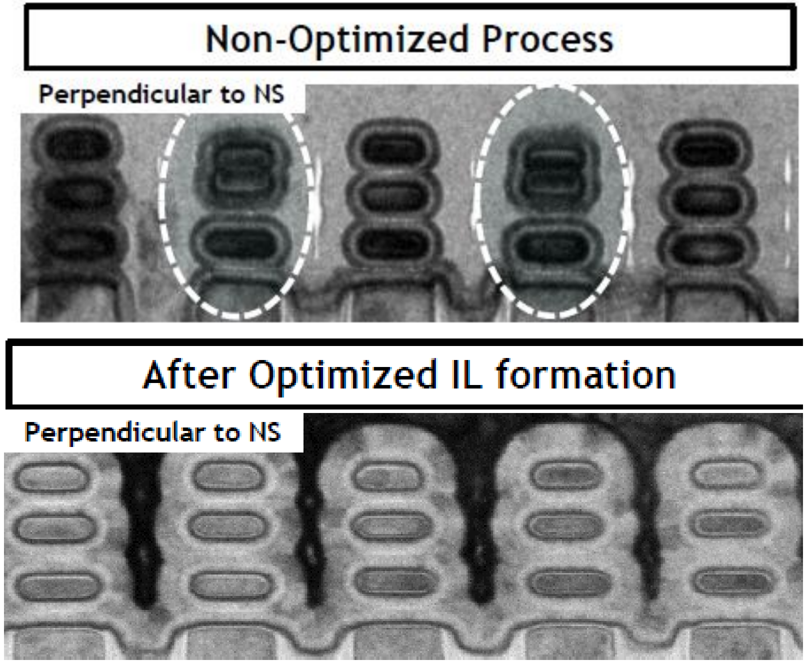


📖 N. Loubet , IBM, VLSI 2020

- Complete removal of SiGe. Selectivity > 100:1 (SiGe:Si)
- Target → 5nm sheet thickness t_{Si} with 0.5nm max variation.
- The thinner the sheet, the stronger the V_t and performance variation due to quantum effects

Nanosheet technology challenges

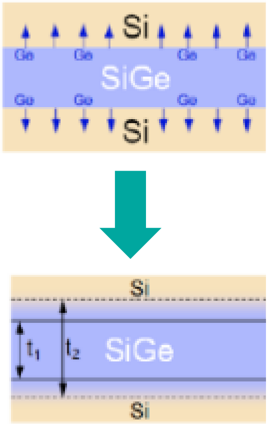
Long-channel stiction

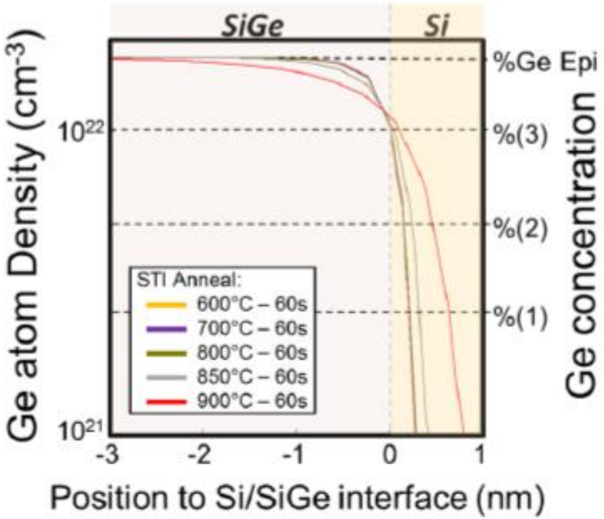


N. Loubet , IBM, VLSI 2020

Limited thermal budget

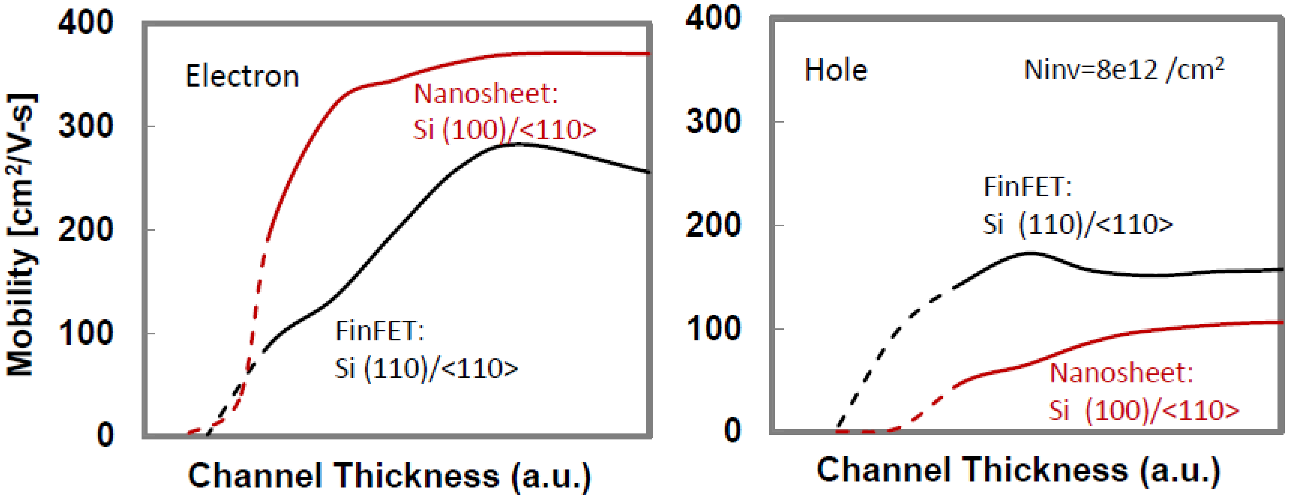
- Ge diffusion in Si → impacts channel release.
- Limited to 1min @850 C ~0.4 nm

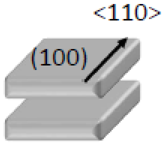




Nanosheet mobility

- Nanosheet channel mostly on (100) surface vs. FinFET on (110)
  - Highly unbalanced mobility: $\mu_N >> \mu_P$





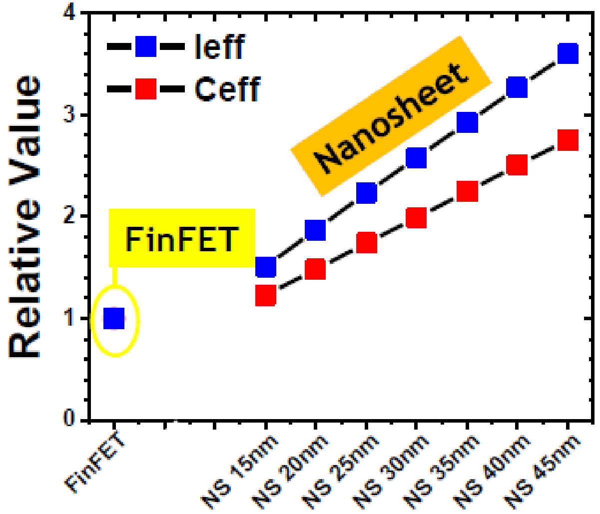
Jin Cai, TSMC IEDM 2019

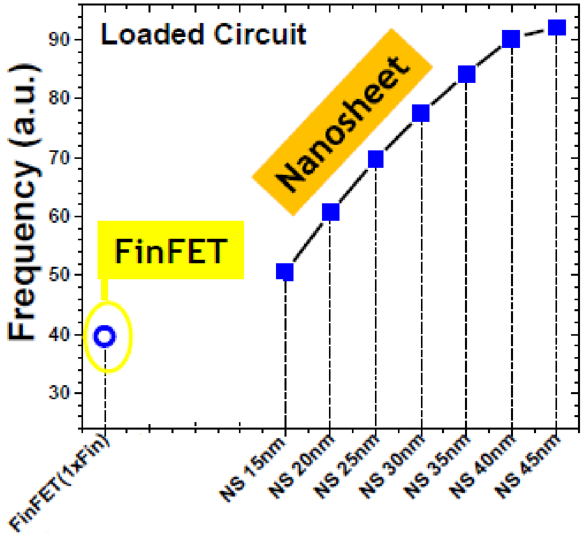
- (110) FinFET vs. (100) Nanosheet: stronger N, weaker P
  - Adjusting W_N/W_P ratio can not fully recover delay penalty

EPFL

Performance Benchmarking of NS vs. FinFET

- Relative speed of FinFET and stacked Nanosheet-FET:

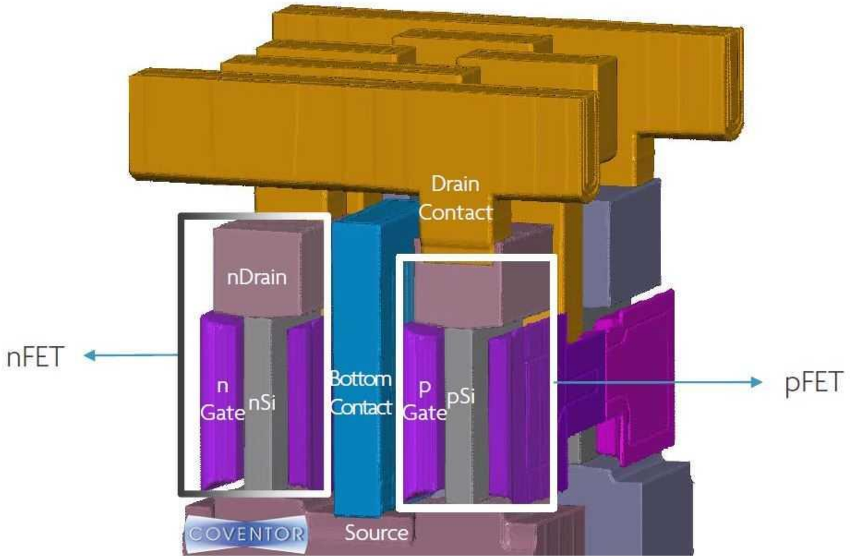




N. Loubet , IBM, VLSI 2020

- Nanosheet-FET has superior intrinsic performance for any sheet width compared to FinFET → better I_{eff}/C_{eff} trade-off
  → Reduced C_{eff} for a given Active width

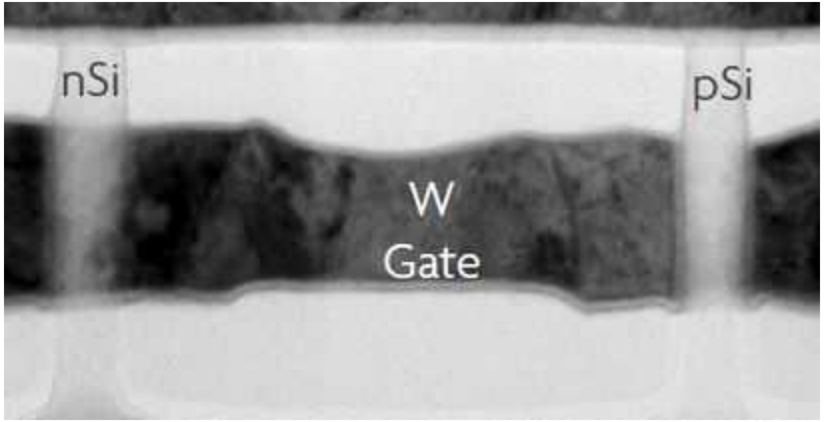
IMEC and academic actors



- Move from 2D to 3D lay-out
- Design and Lay-out very different
- gate length can be more relaxed without consuming a larger area on the wafer
- Relaxation in the nanowire diameter while preserving control over SCE → opportunity for high-mobility channels (limited by quantum effects in scaled channels)

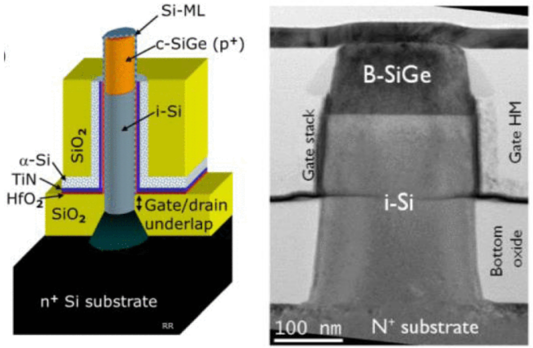
📖 https://www.imec-int.com/en/imec-magazine/imec-magazine-september-2017/the-vertical-nanowire-fet-enabler-of-highly-dense-srams

Device integration: the channel-first approach

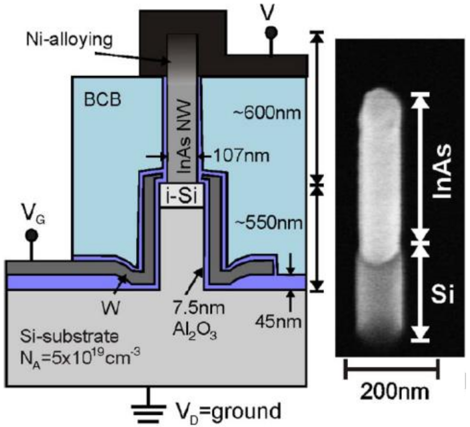


- Nanowire pillars are formed and doping is introduced prior to other processing steps.
- Different dopant concentrations for a nMOS/pMOS.
- Junction-less is another option

- Opportunities for alternative channel/Source materials

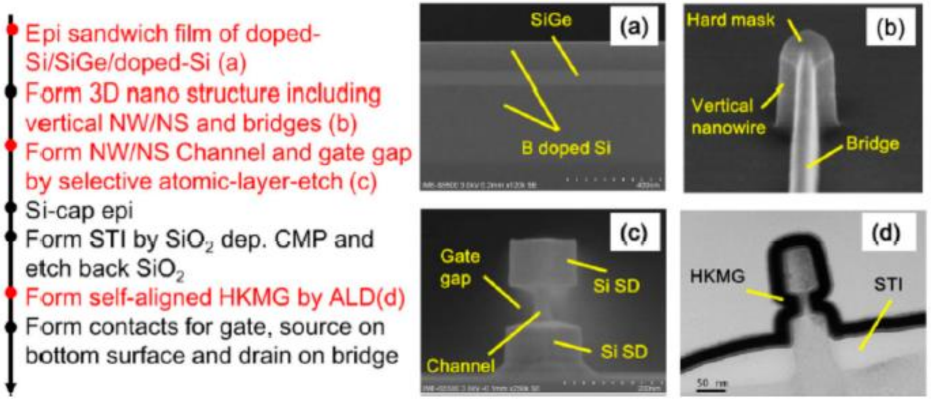


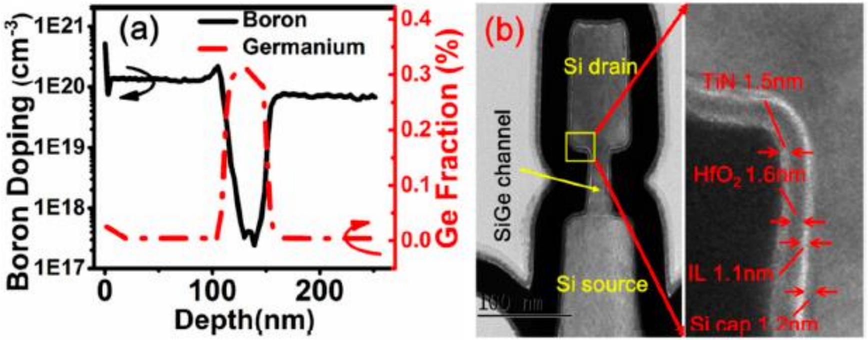
📖 P. G. D. Agopian, TED 2015



📖 K. Moselund, EDL 2012

IMEC work on Vertical GAA FET





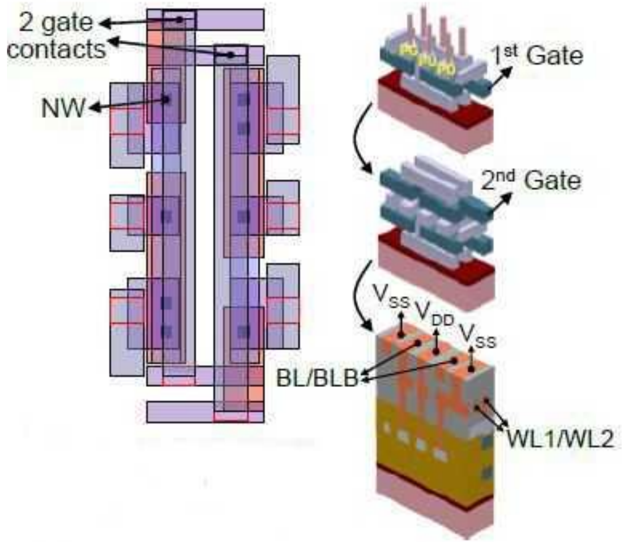
- Epitaxial sandwich structure of SiGe as vertical channel → Well defined L_G
- V-GAAFET t_{Si} = 20 nm, L_G = 60 nm
- SS_{sat}= 86 mV/dec, I_{on}= 37.6 uA/um, V_d = 0.65 V
- Projection w. buried interconnect:
  - Area reduction - 22.5%,
  - Wire length reduction 14.4%
  - Capacitance reduction 28.4%

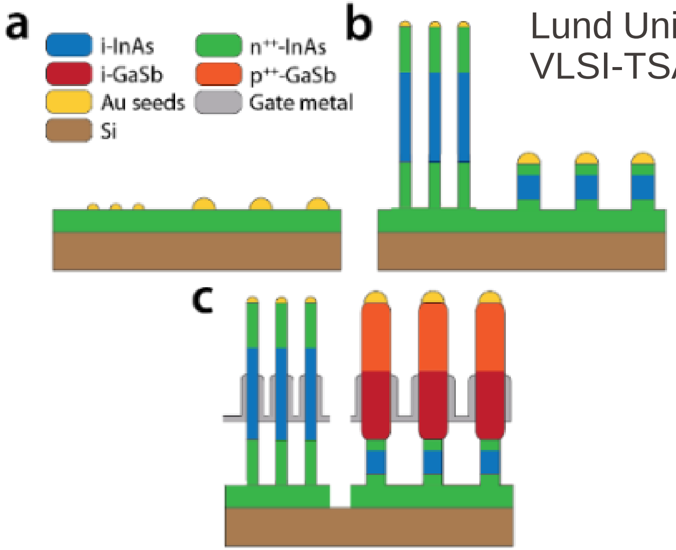
📖 X. Jin et al, EDL 2019

Going vertical

Dense design without performance degradation, concepts

📖 L.E. Wernersson, Lund University, VLSI-TSA 2017





- Vertical stacked 6T SRAM proposed by IMEC
- 39% area reduction (per bit)

- Complementary III-V process, vertical direction exploiting different growth modes.

Going vertical is not new – 3D NAND flash

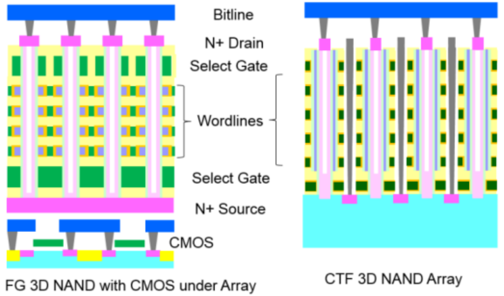


Fig. 3. The two primary 3D NAND technologies - FG Flash 3D NAND with CMOS under array & CTF 3D NAND.

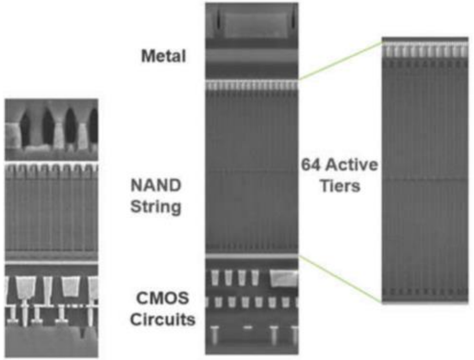
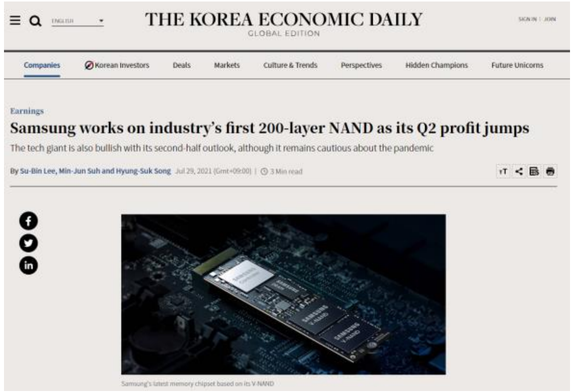


Fig. 4. SEM cross-sections of the 1st and 2nd Gen FG 3D NAND with 32 and 64 active Layers and CMOS under the array.

📖 K. Parat, Intel, IEDM 2018


THE KOREA ECONOMIC DAILY
GLOBAL EDITION

Companies | Korean Investors | Deals | Markets | Culture & Trends | Perspectives | Hidden Champions | Future Unicorns

Earnings

Samsung works on industry's first 200-layer NAND as its Q2 profit jumps

The tech giant is also bullish with its second-half outlook, although it remains cautious about the pandemic

By Su-Bin Lee, Min-Jun Suh and Hyung-Suk Song Jul 29, 2021 (Gmt+09:00) | 3 Min read

Samsung's latest memory chipset based on its V-NAND

- For Memory, stacking is crucial to achieve integration density → interesting to observe microelectronics tech. development
- 2020 SK Hynix unveiled 176 layer NAND
- 2022 Samsung working on 200+ layer stack

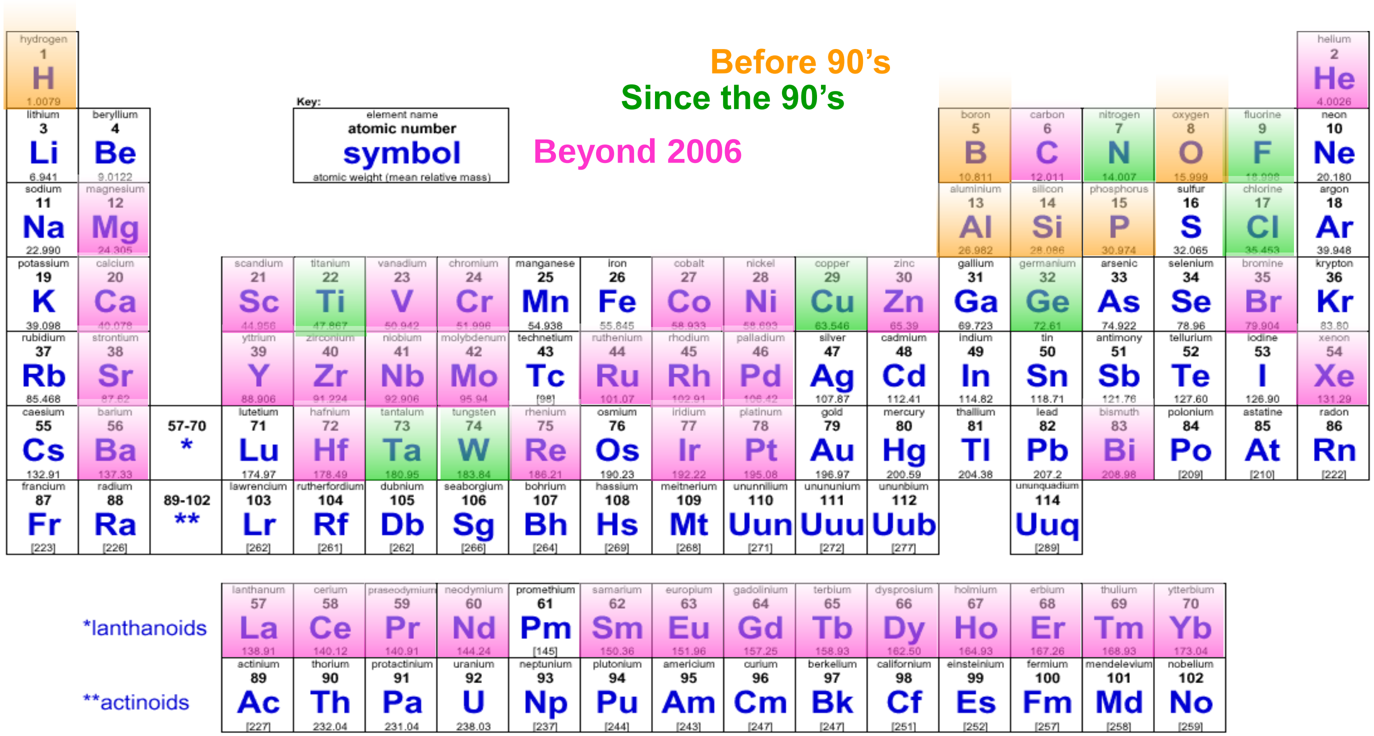
Device architectures summary

Options	Challenges	Strength	Readiness
Planar	SCE, limited scaling	Simple, low-cost, reliable	Production
FDSOI	SOI substrate	Low-power, better SCE than planar	Production
FinFET	Fin width scaling, limited Fin height	N/P balance	Production
Nanosheet	N/P imbalance, max. number of stacks	Improved SCE, sheet-width scalable	Development
Vertical NW	Lay-out & fabrication, RC	3D → relaxed dimensions	Research

CMOS Fabrication

Improving CMOS Performance: Materials Innovations

Elements Employed in Silicon Technology



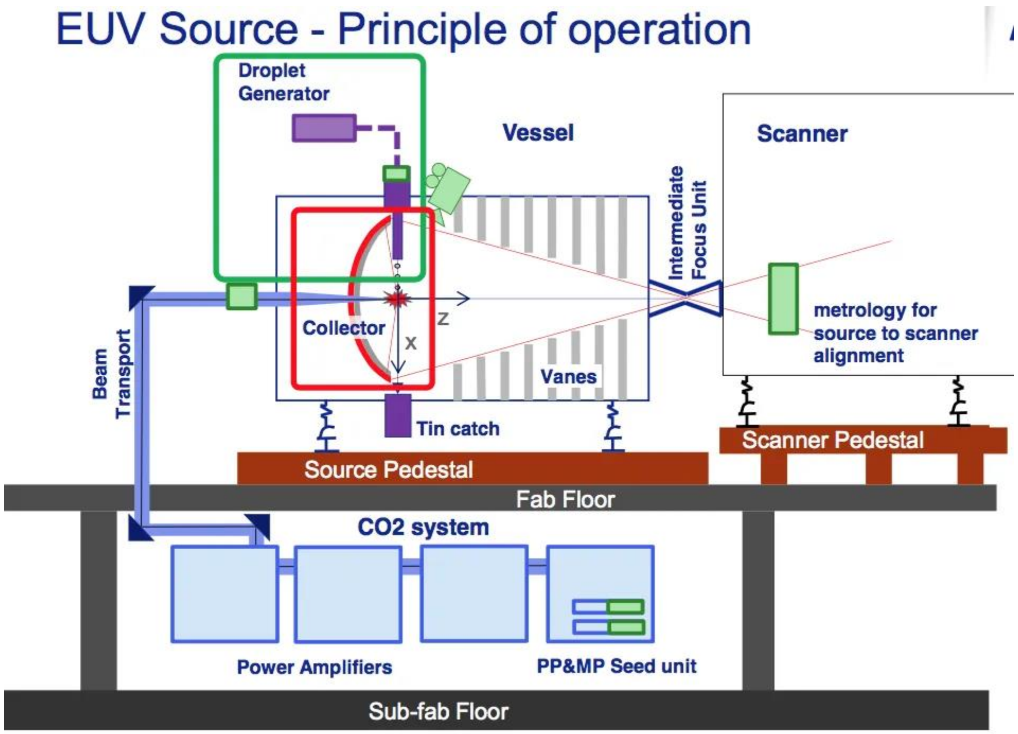
➔ Material Innovations ➔ Complexity ➔ Cost $$$!

Advanced lithography process

- **FinFETs 20-nm and 14-nm node technologies:** 193 nm ArF immersion litho with multiple patterns has been mainly used in manufacturing
- **For 7nm and beyond:** 193 nm immersion with self-aligned double pattern (SADP) and self-aligned quadruple pattern (SAQP).
  - 193nm lithography reached its limit at 80nm. But chipmakers extended 193nm lithography far below this wavelength by using resolution enhancement techniques
- **Extreme ultraviolet (EUV) lithography** has been recognized as a promising candidate for the manufacturing of semiconductor devices as line space (LS) and contact hole (CH) patterns for 7 nm node and beyond.
- Single-patterning process production cost is greatly lower than that of “multi-patterning” of repeated pattern circuits → cost benefit of EUV.

EUV - the most complex piece of machinery in the history of the IC industry

- A power source converts plasma into light at 13.5nm wavelengths
- Droplet generator produces 25 μm tin droplets
- A CO_2 laser to fires a pre-pulse → turns the droplet into a pancake-like shape.
- The main pulse hits the pancake-like tin droplet and vaporizes it creating a plasma which emits EUV light at 13.5 nm
- Need to hit each droplet twice with the pre-pulse and main-pulse at 50,000 times a second.

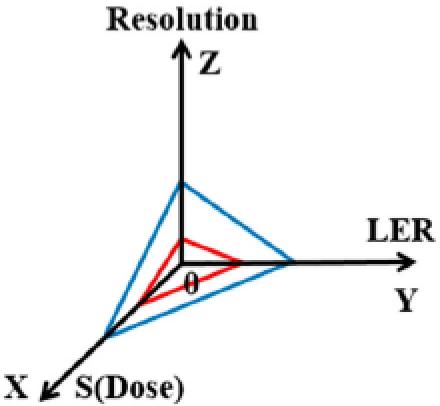


https://semiengineering.com/why-euv-is-so-difficult/

EUV systems

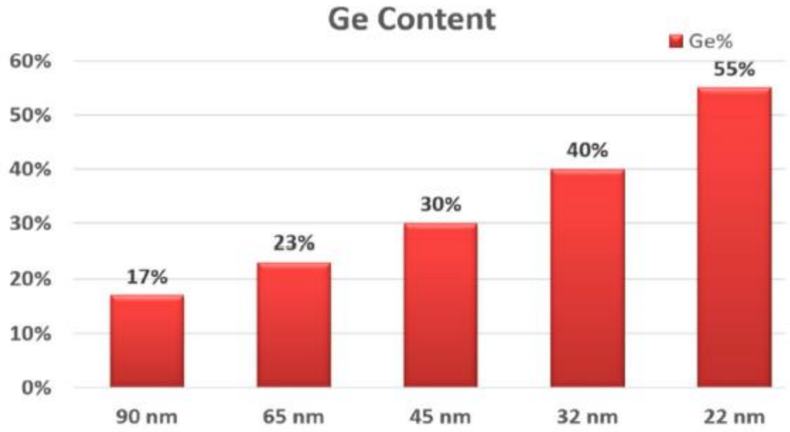
- Light source power is critical for high-volume manufacture (HVM). High throughput = cost advantage
- Current standard ASML NXE3400B (250W EUV) → enables wafer throughput up to 140 wph.
- The masks operate in reflective mode
- Challenge EUV resists: correlation of resolution, line edge roughness, and sensitivity.

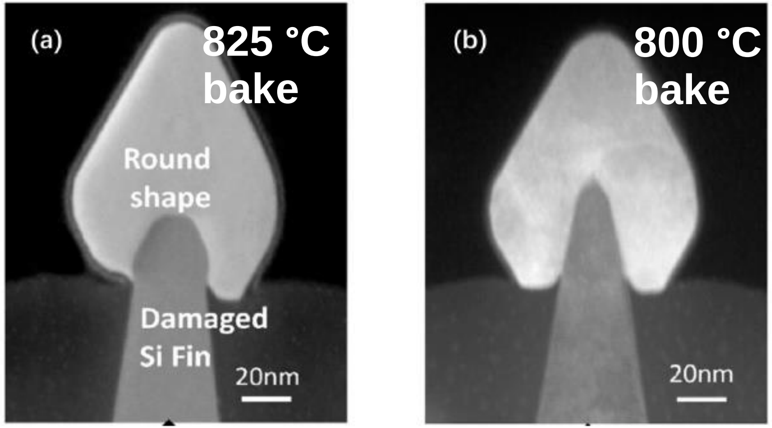
📖 Henry H. Radamson et al. IMEC, Nanomaterials 2020



Strain as mobility booster

- Strain is used as mobility booster in Si-based processes
- Different strain beneficial for n- and p-channel devices
- Methods of strain application
  - Gate stack technology
  - Ge S/D processs. Increasing Ge content for smaller nodes
  - Shallow trench isolation
  - Nitride Contact-Etch Stop Layer (CESL)
- SEG of SiGe for S/D: critical process for device perfromance. Pattern and shape dependent. Critical T-dependence.

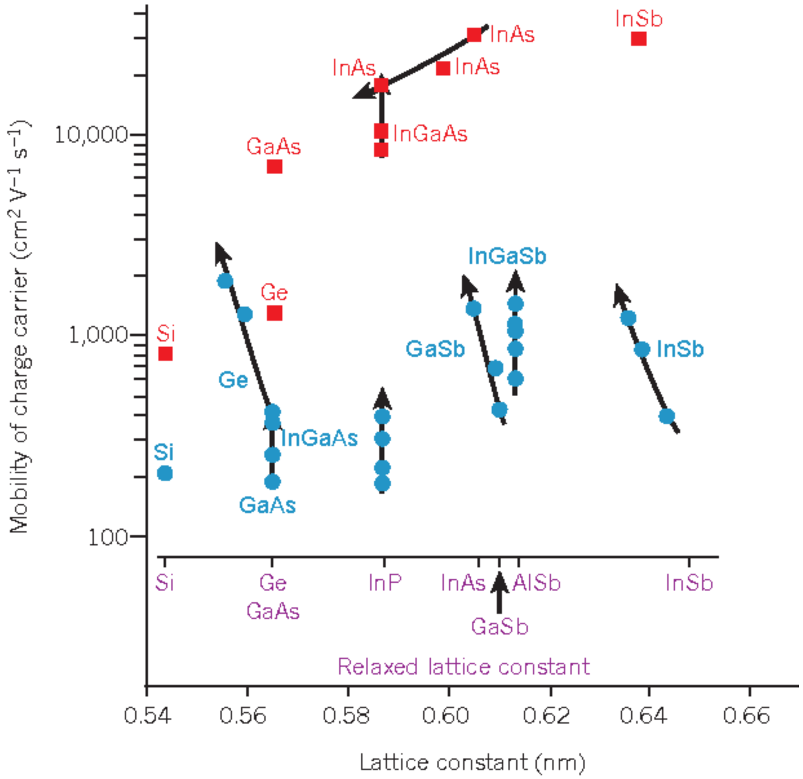




Henry H. Radamson et al. IMEC, Nanomaterials 2020

High-mobility channels

- Mobility in Si is limited, especially for holes.
- Planar processes – can be more easily balanced by width-scaling
- Strain is used extensively to boost mobility in Si-technologies



J. del Alamo, Nature 2011

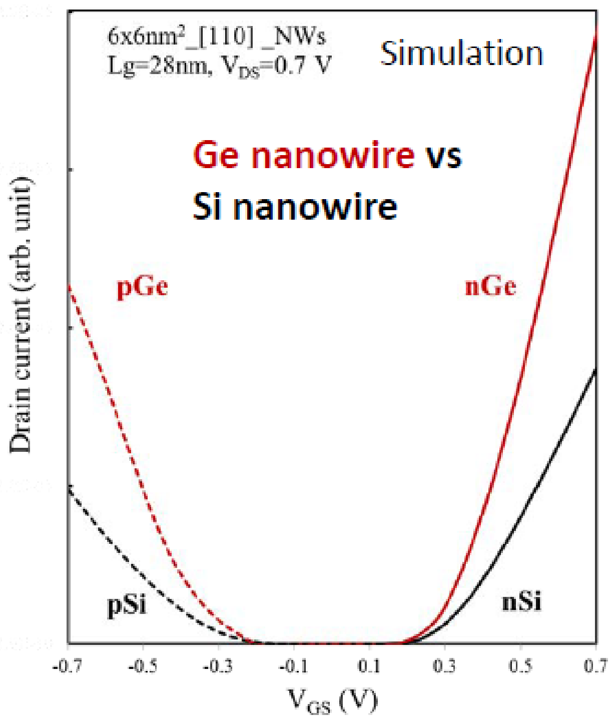
Ge Mobility and reduced leakage in NWs

- Ge has high electron/hole mobility and high DOS
- Confined channel geometry increases band gap and effective mass and suppresses BTBT leakage

NW: 6nm x 6nm nanowire

	Si (bulk)	Ge (bulk)	Si NW	Ge NW
μ_N (cm2/V-s)	1600	3900	300[1)]	900[1)]
m_t/m_l	0.19/0.916	0.082/1.46	0.295[2)]	0.092[2)]
μ_P (cm2/V-s)	430	1900	110[1)]	255[1)]
m_{HH}/m_{LH}	0.49/0.16	0.28/0.044	0.446[2)]	0.125[2)]
E_G (eV)	1.12	0.66	~1.3	~1

1) Mobility at 5e12/cm^2 2) Transport mass at band edge



A. Toriumi, T. Nishimura, 2018 Jap. J. Appl. Phys. (U Tokyo)
S.-K. Su, E. Chen, J. Wu, 2018 SISPAD (TSMC)

📖 Jin Cai, TSMC
IEDM 2019

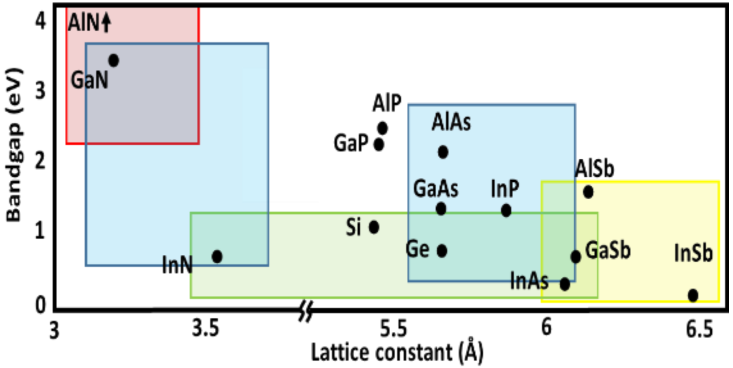
New Materials: Merits of III-V

Characteristics

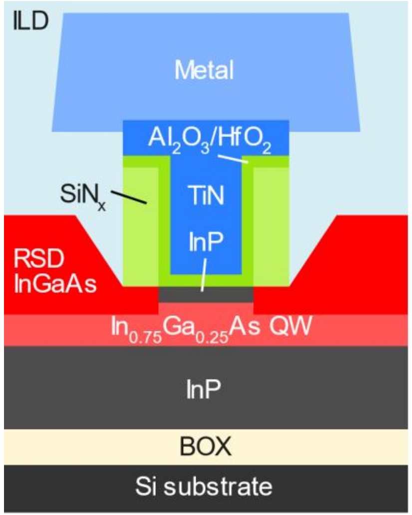
- High electron mobilities
- Optically active
- Tunable bandgap by composition
- Quantum phenomena start at larger dimensions (lower DOS)
- Large lattice and thermal mismatch to silicon → integration is challenging

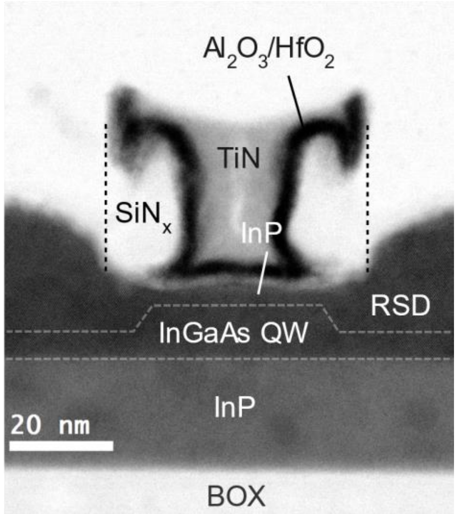
Material/Property	Si	Ge	GaAs	InAs	InSb
m_{eff}^*	0.19	0.08	0.067	***0.023***	***0.014***
μ_n (cm^2/Vs)	1600	3900	9200	***40,000***	***77,000***
E_G (eV)	1.12	0.66	1.42	***0.36***	***0.17***
ε_r	11.8	16	12.4	14.8	***17.7***

A. Pethe, et al. IEDM 26.3, 2005

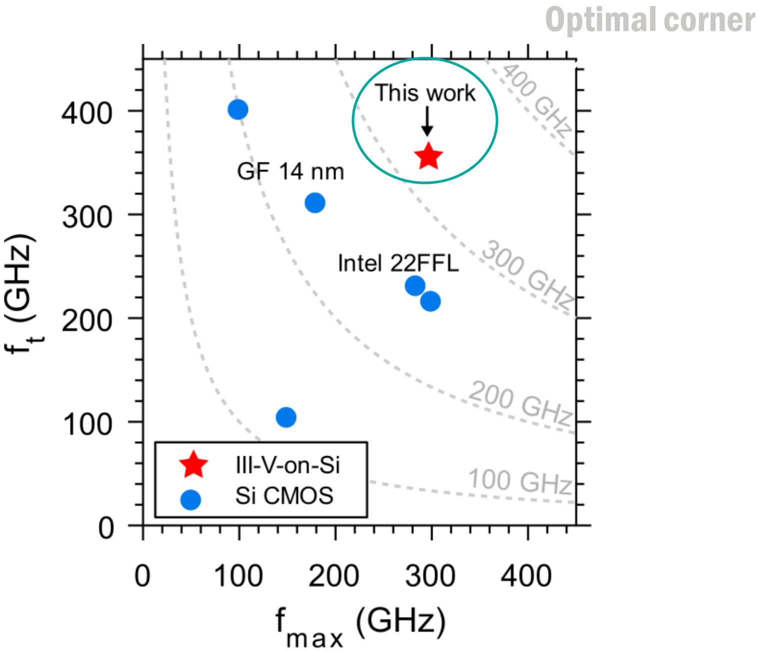


III-V RF technology platform





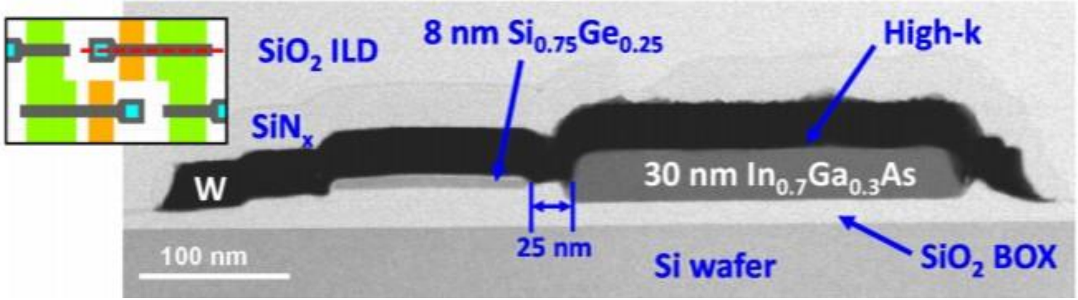
C. B. Zota et al., *Electron device letters*, 2018

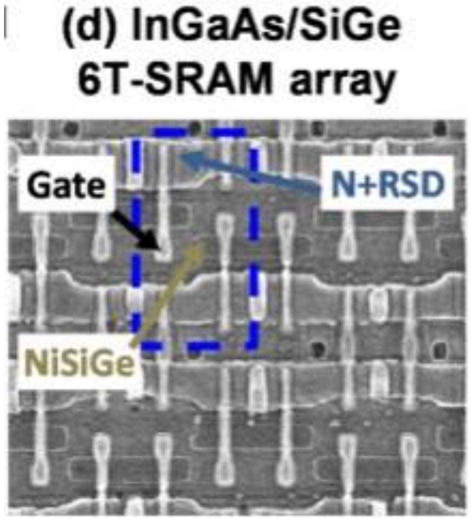


- **CMOS-compatible** and self-aligned fabrication flow
- Based on in-house wafer bonding. Development of III-V process blocks
- State-of-the-art RF performance on silicon

Demonstration of advanced InGaAs FETs

Hybrid III-V/SiGe CMOS



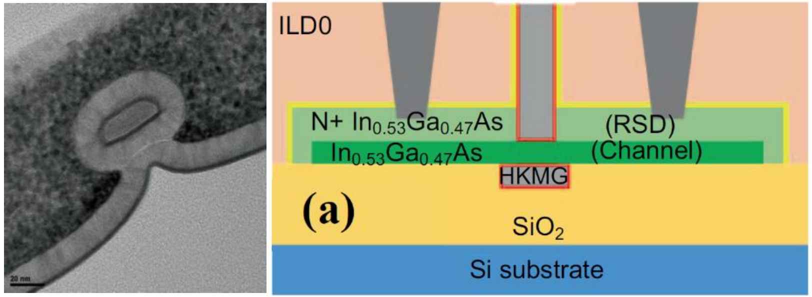


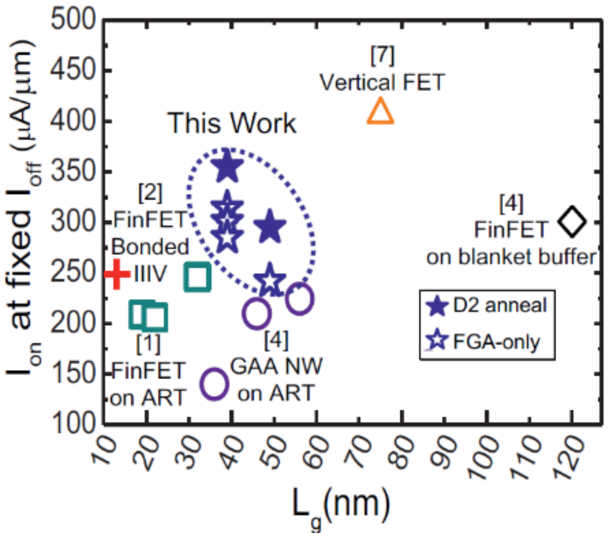
L. Czornomaz et al. VLSI (2015)

- The first and only demonstrations of InGaAs/SiGe circuits on Si are from IBM (ZRL)
- Based on TASE growth of III-V virtual substrates
- Combination with advanced RF process technology

Transfer to advanced CMOS platform

- III-V nanosheet MOSFETs on Si
- Bringing III-V channel FETs into advanced CMOS platform
- Development of growth and process modules at ZRL and transfer of technology to MRL 200 mm process line (IBM Yorktown)

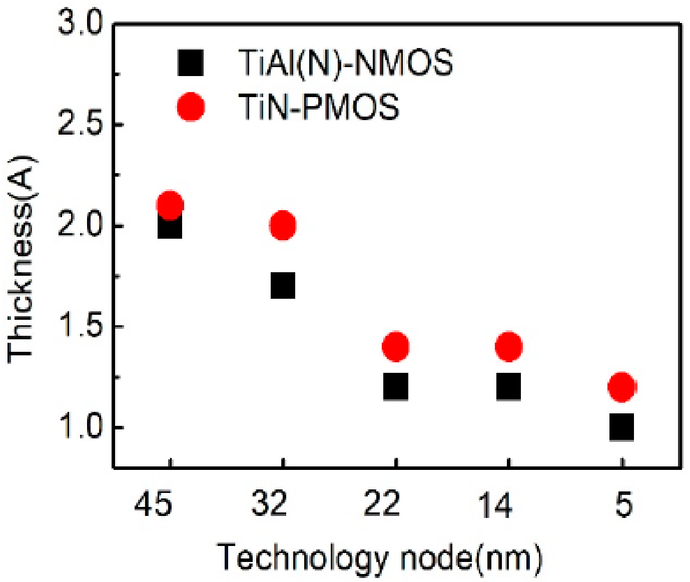




📖 S. Lee et al. IEDM 2018 (highlight paper)

Gate Workfunction Metalization

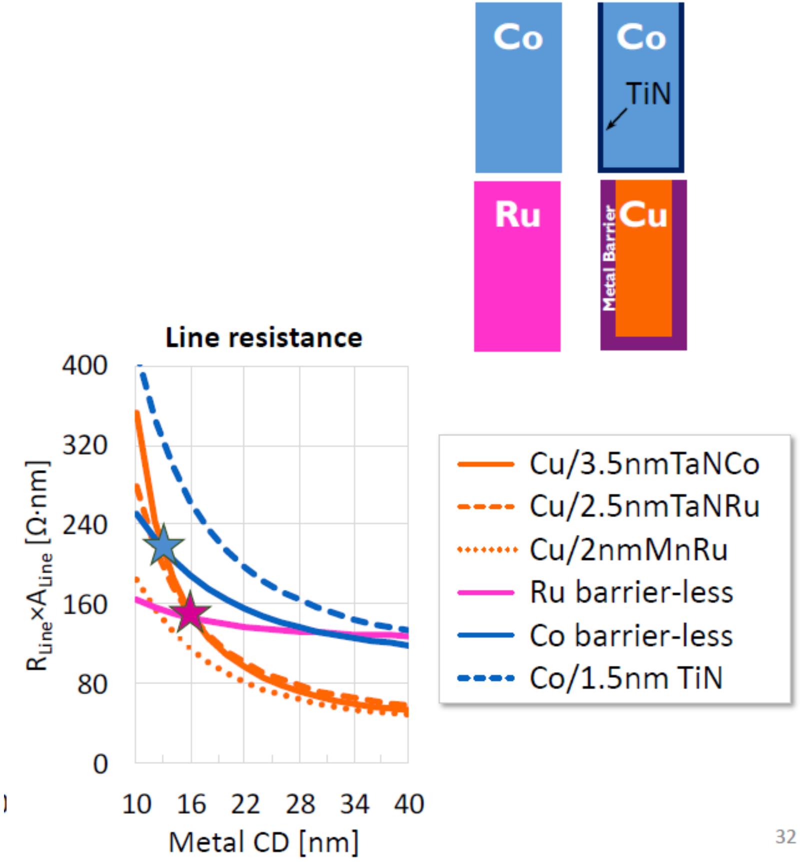
- Different metals for workfunction engineering
  - NMOS: TiAl(N)
  - PMOS: TiN
- Filling up with W
- Nanosheet technology: need ALD metals for coverage = challenging to control workfunction



📖 Henry H. Radamson et al. IMEC, Nanomaterials 2020

Contact Metalization

- Cu interconnect beyond 7 nm is difficult, because of scalability of TaN liner.
- TaN ALD might be an option
- Barrier layers based on Co/Ru is an option
- For the same trench cross-section, barrier-less Ru and Co can outperform Cu
  - For Ru, the first cross-over with Cu occurs at around 16nm
  - For Co, the first cross-over with Cu occurs at around 12nm



32

I. Ciofi (imec). TED 2019

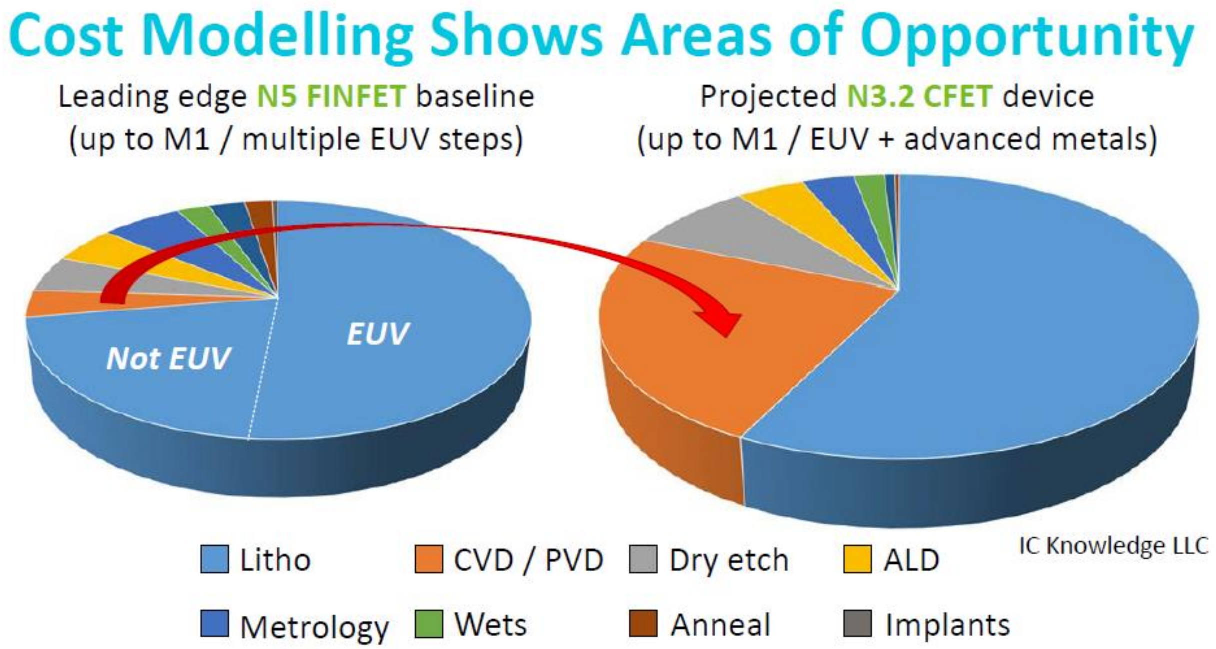
Advanced nodes require more rare metals

Rare World Metals Mint

- Smaller features require higher conductivity metals such as Co and Ru
- Price of advanced metals seems to scale with thin-film conductivity
- Metal consumption key part of cost reduction

CMOS production cost

Different cost segments dominate depending on device architecture and technology node.



L. Liebmann, TEL, IEDM 2019

Partial Summary: CMOS processing

- EUV has been essential in pushing single-digit nodes, but it adds a lot of complexity and cost
- Silicon mobility is limiting especially for 3D architectures like nanosheets. Different materials (SiGe and III-Vs) might help to boost mobility.
- Technology challenges includes
  - Controlling strain and mobility in stacked nanosheet technologies
  - Contact and line resistance
  - Gate dielectric scaling
- Total cost and cost segmentation evolves for different nodes and device architectures

Summary

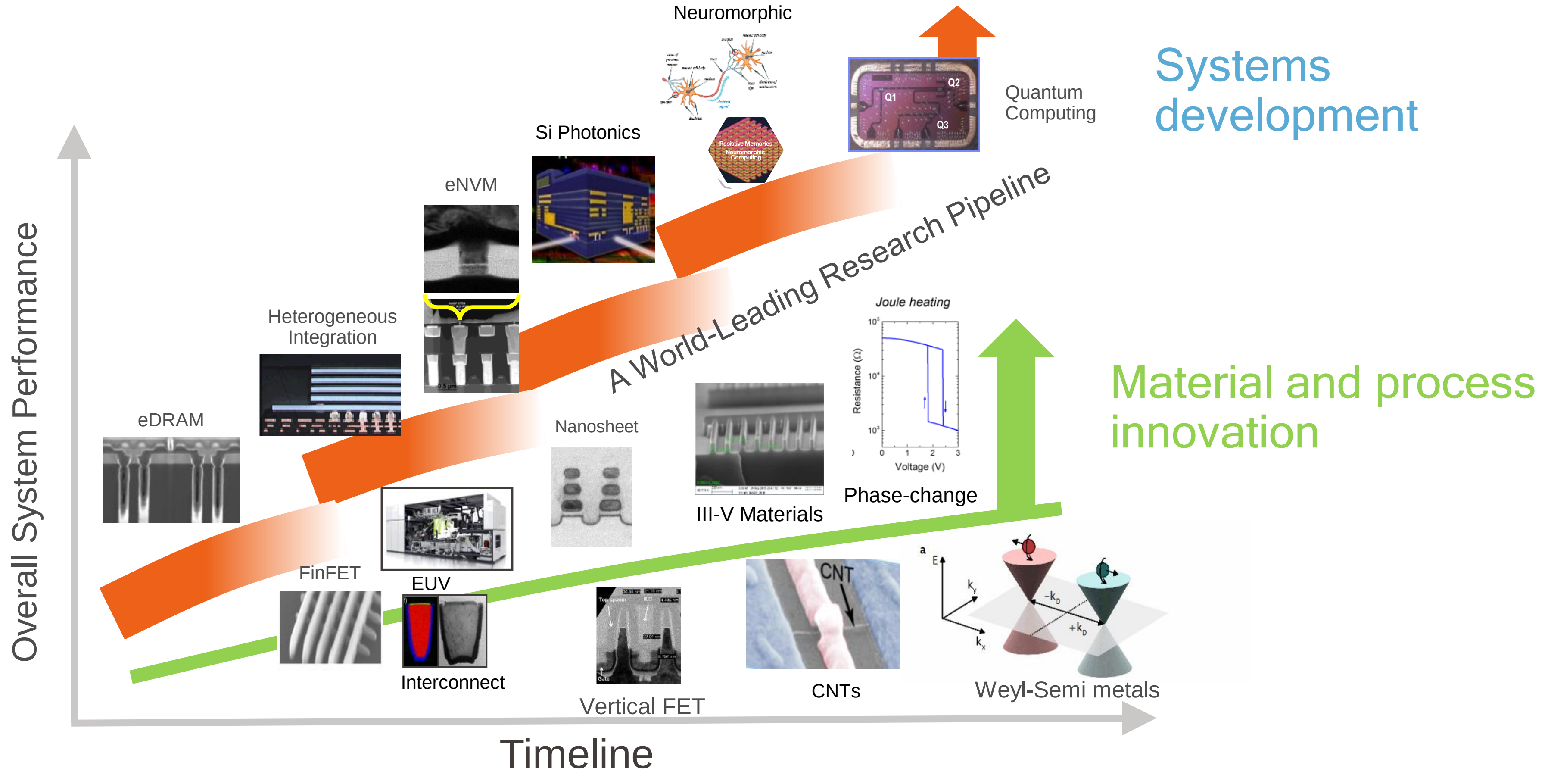
Summary

- Moore's law in terms of device density and functionality continues.
- A lot of innovation is required to maintain performance and pure scaling is no longer enough
- For 14nm nodes and beyond advanced architectures like FinFETs and stacked nanosheets are required
- This entails massive challenges in process optimization and development of new tools for 3D technology processes

Here we covered only CMOS logic scaling.
In terms of the future of computing entirely new domains open up within AI and quantum computing. Increasing focus on memory.
→ There's never bin a more interesting era for microelectronics

Systems Performance Roadmap

Through Differentiated Technology



Thank you for your attention – Questions?

PSI team 2022

IBM team 2020

Funding
EU FP7: E2SWITCH
EU H2020:CONNECT, DIMENSION, INSIGHT, DESIGN-EID

See you, back on Friday for integrated photonics