

Fundamentals of Analog & Mixed Signal VLSI Design

Technology Roadmap

Christian Enz

Institute of Electrical and Micro-Engineering (IEM), School of Engineering (STI)

Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland

The logo of the Swiss Federal Institute of Technology, Lausanne (EPFL), consisting of the letters 'EPFL' in a bold, red, sans-serif font.

Outline

- **Introduction**
- CMOS technology scaling
- Power consumption and energy efficiency
- Voltage scaling

Chips Have Changed our Daily Life...

Data Center & Cloud



PC

2 in 1



Mobility

TABLET



Wearable

GLASSES



IoT

SMART CITIES



STORAGE



DESKTOP



SMARTPHONE



PERSONAL ASSISTANT



SMART AGRICULTURE



NETWORKING



ALL IN ONE



PHABLET



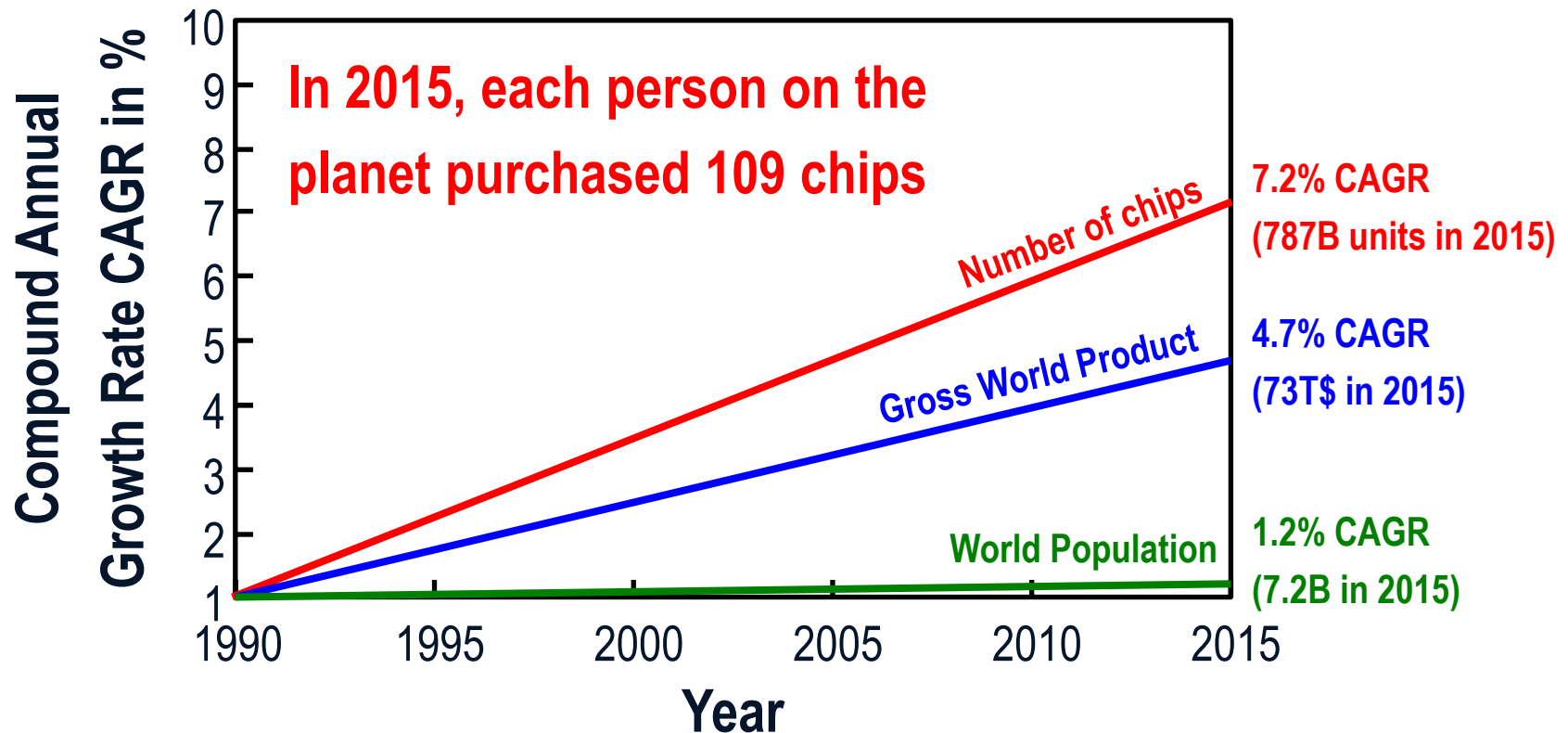
SPORTS FASHION



SMART FACTORIES



Semiconductor Unit Growth

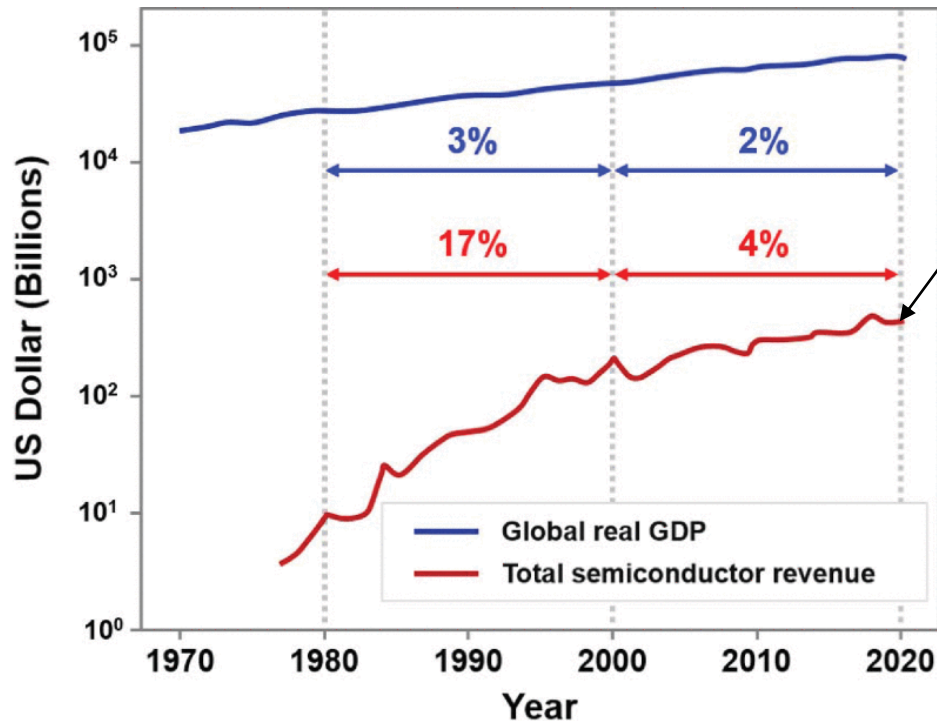


- Note that the **compound annual growth rate** or CAGR of a variable $V(t)$ is defined as follows

$$CAGR(t_0, t_n) = \left(\frac{V(t_n)}{V(t_0)} \right)^{\frac{1}{t_n - t_0}} - 1$$

- where $V(t_0)$ is the start value, $V(t_n)$ the final value and $t_n - t_0$ the duration in years

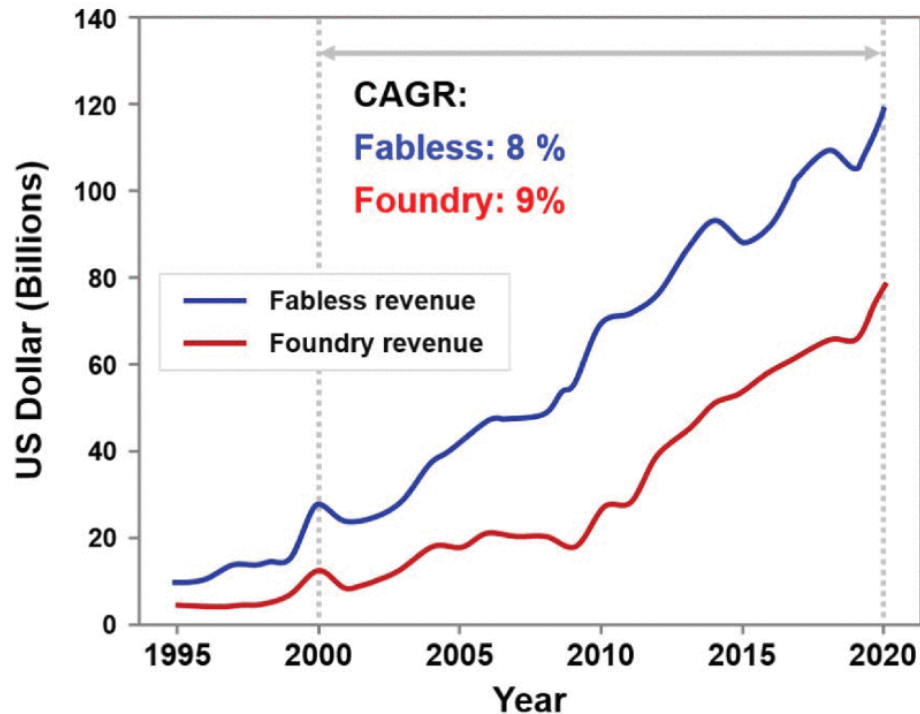
Growth of the Semiconductor Business



The global semiconductor market is estimated at 450 billion USD in revenue for 2020

- The growth rate of revenue of semiconductors parallels those of the gross world product (GWP) for the past 20 years
- After the initial fast growth period around the 1990s, worldwide semiconductor sales grow at a similar rate as the gross world product

Growth of the Fabless and Foundry Business



- From 2000 to 2020, the overall semiconductor industry grew at a steady 4% annual growth rate, the fabless sector continued strong growth at 8%, with the foundry sector at 9%, compared to 2% for the integrated device manufacturers (IDMs)
- Fabless revenue accounts for 35% of total 2020 semiconductor industry revenue, excluding memory, versus 17% in 2000

Outline

- Introduction
- **CMOS technology scaling**
- Power consumption and energy efficiency
- Voltage scaling

Moore's Law

Electronics, April 18, 1965

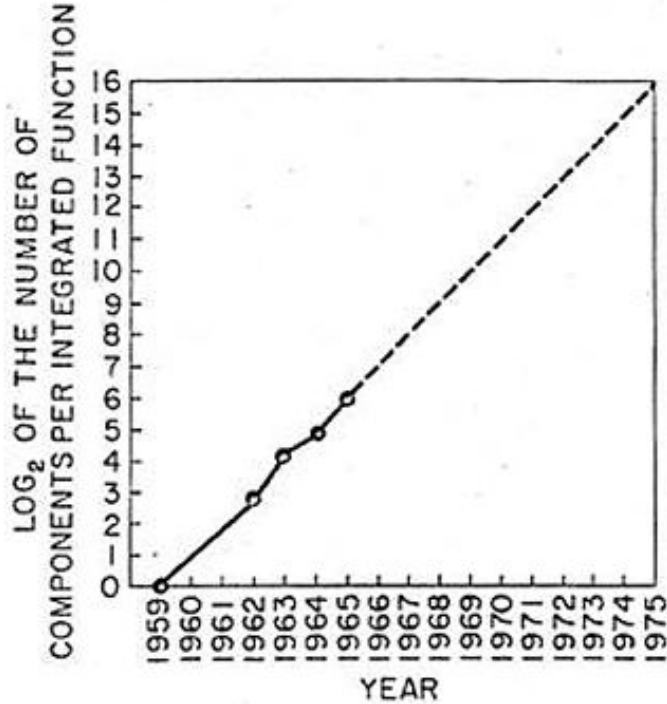
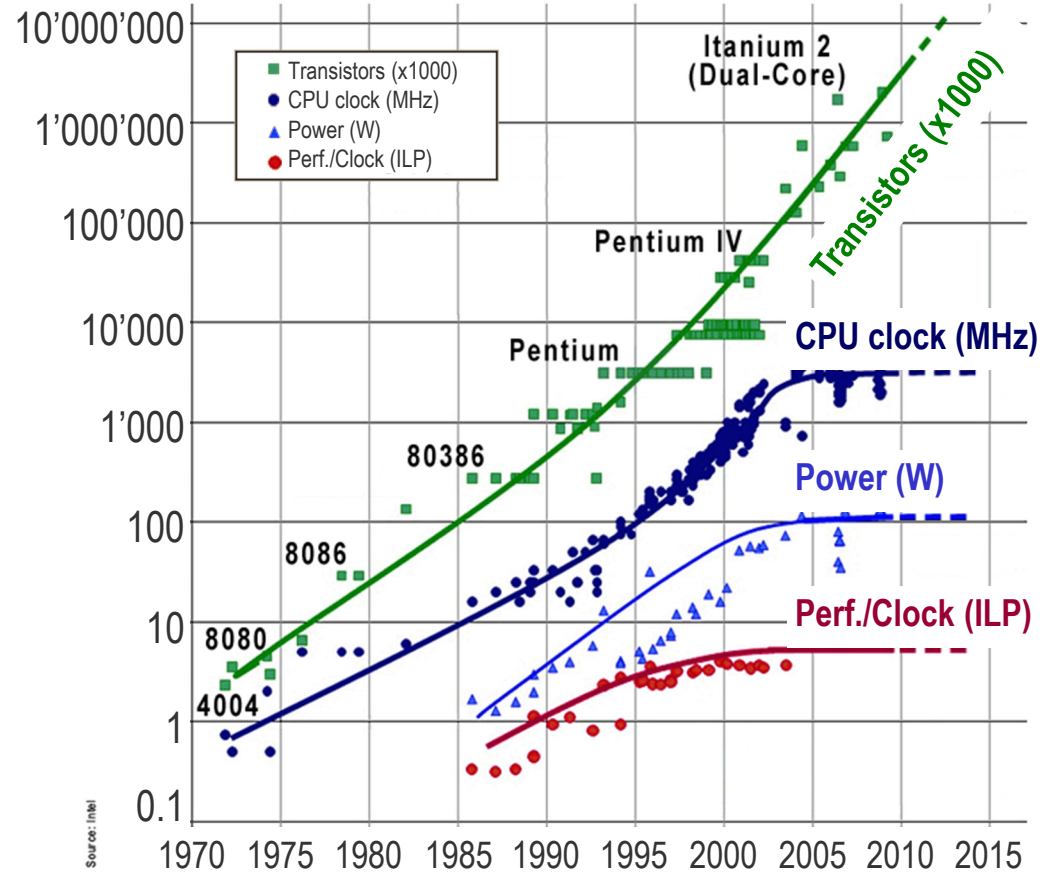
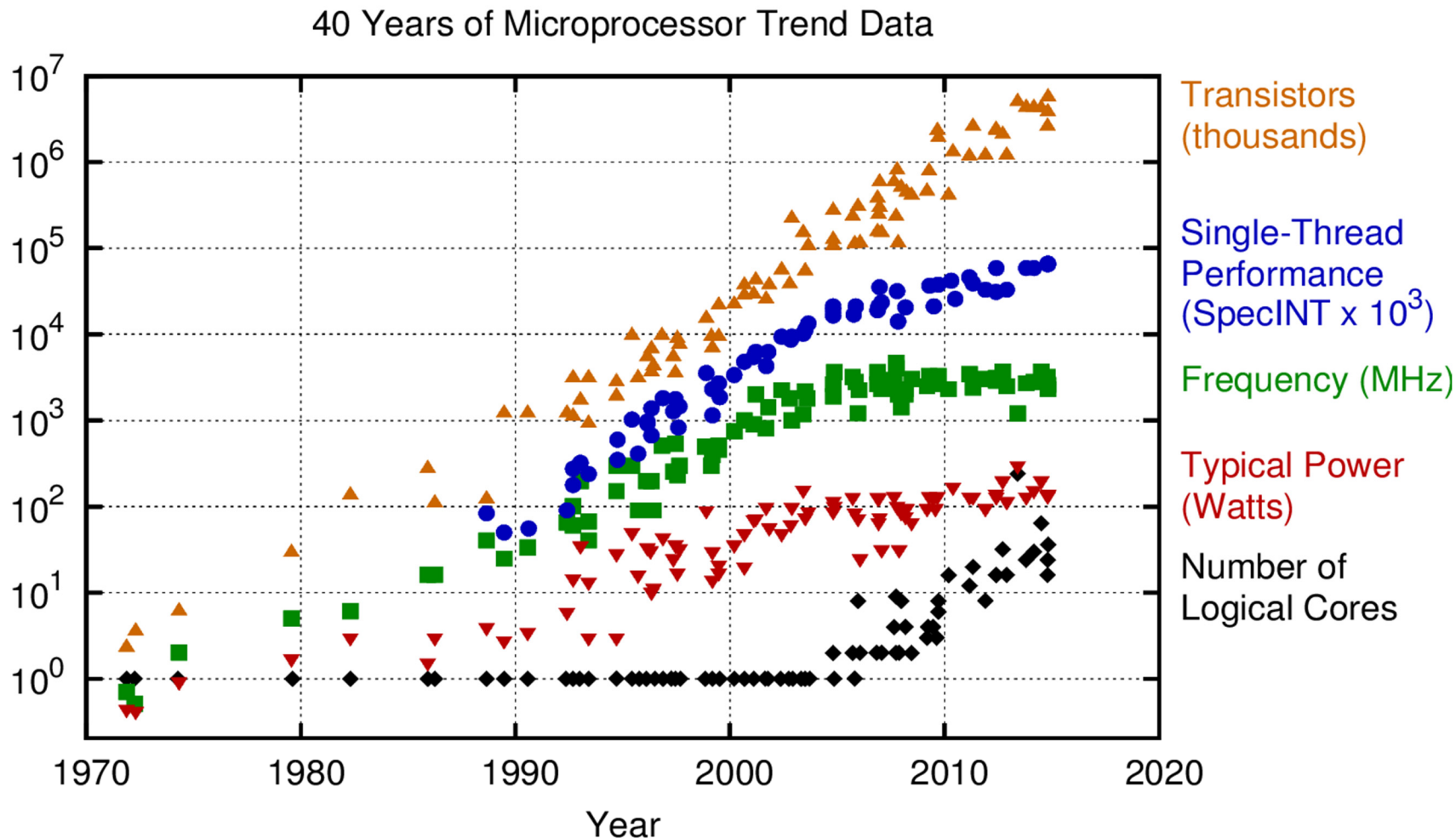


Fig. 2 Number of components per integrated function for minimum cost per component extrapolated vs time.



The number of transistors on a silicon chip **doubles** every **two years**

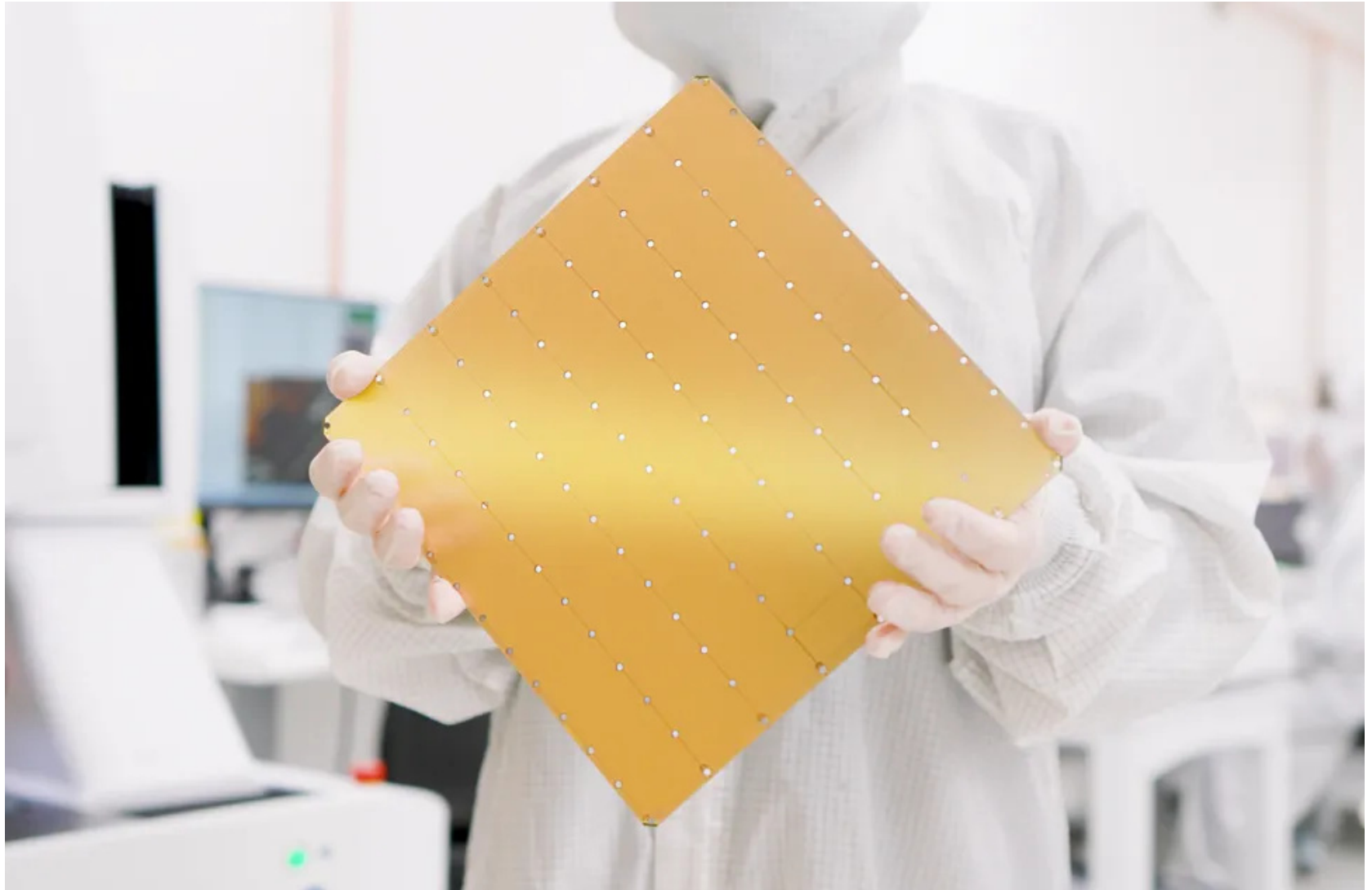
40 Years of Microprocessor Trend Data



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
 New plot and data collected for 2010-2015 by K. Rupp

Source: K Rupp.

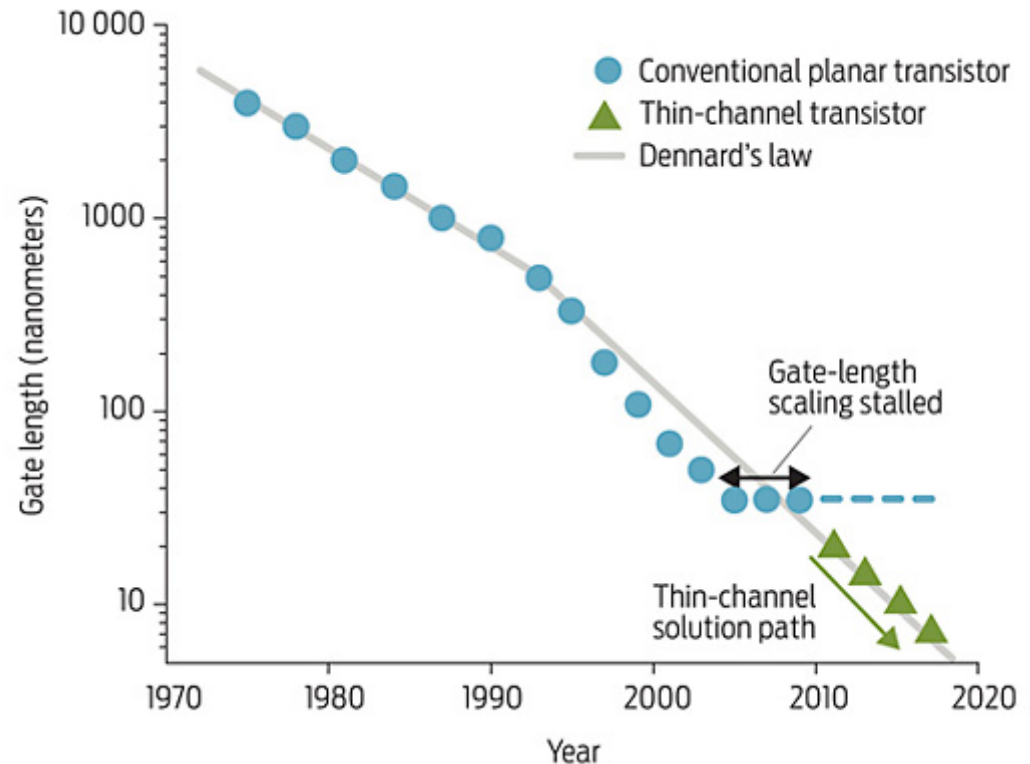
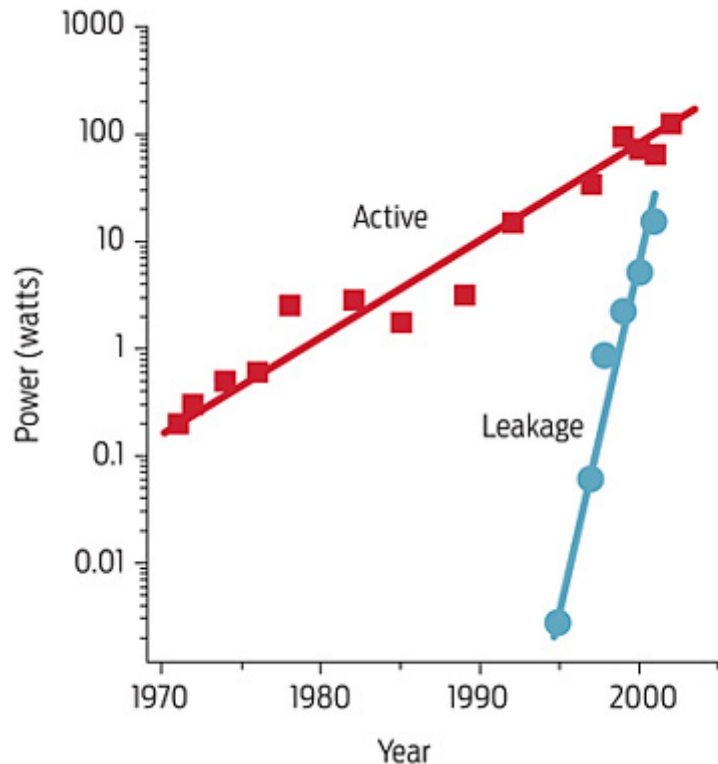
Cerebras' 4 Trillion Transistors Waferscale AI Chip



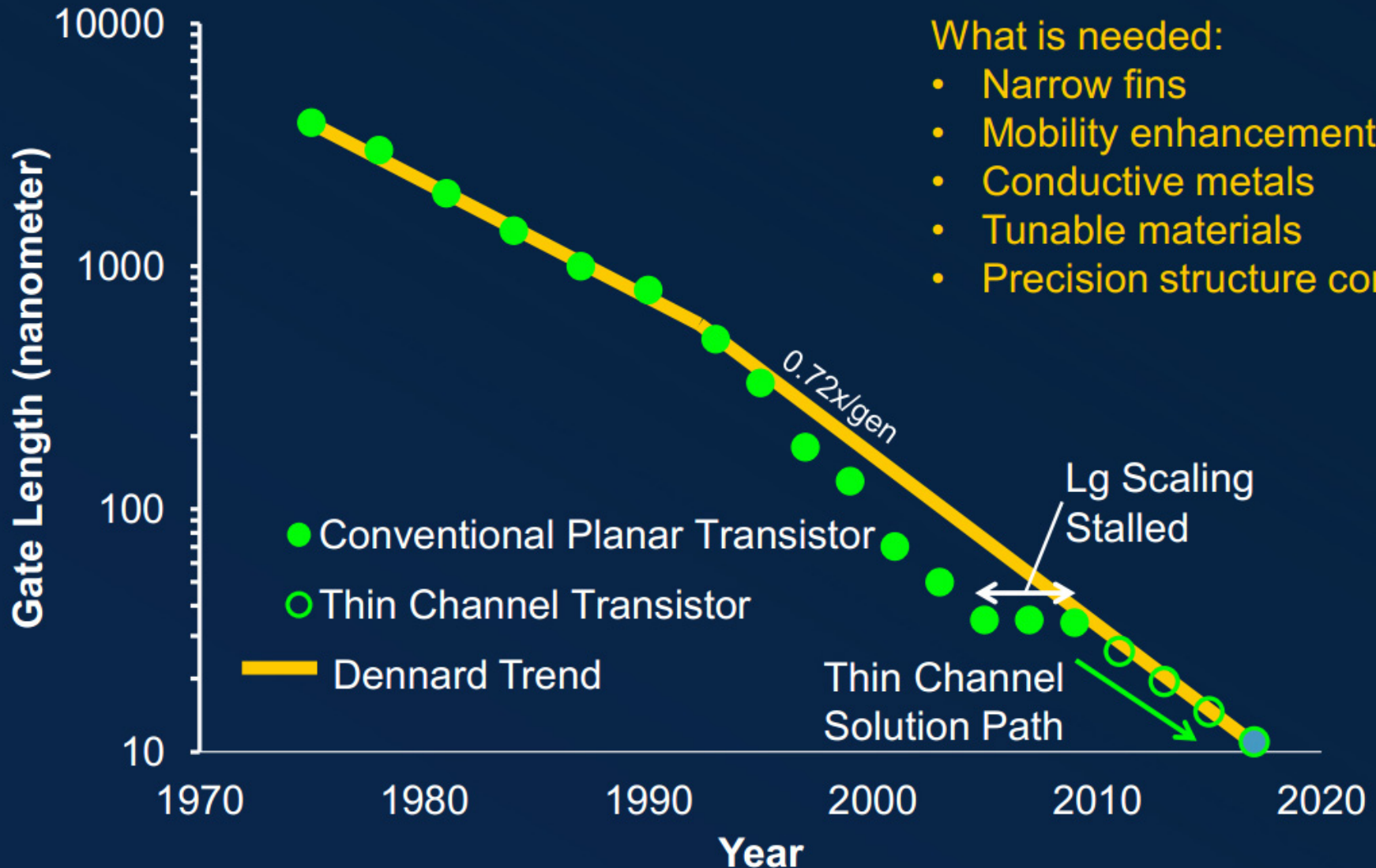
 S. K. Moore, "Cerebras Unveils Its Next Waferscale AI Chip," IEEE Spectrum, March 2024.

CMOS Technology Scaling

- **Leakage power** has caught up with dynamic (active) power and has put a halt to the conventional Dennard's transistor scaling progression
- Switching to **alternate architectures** is required to shrink transistors further, boosting density and performance



CMOS Technology Scaling



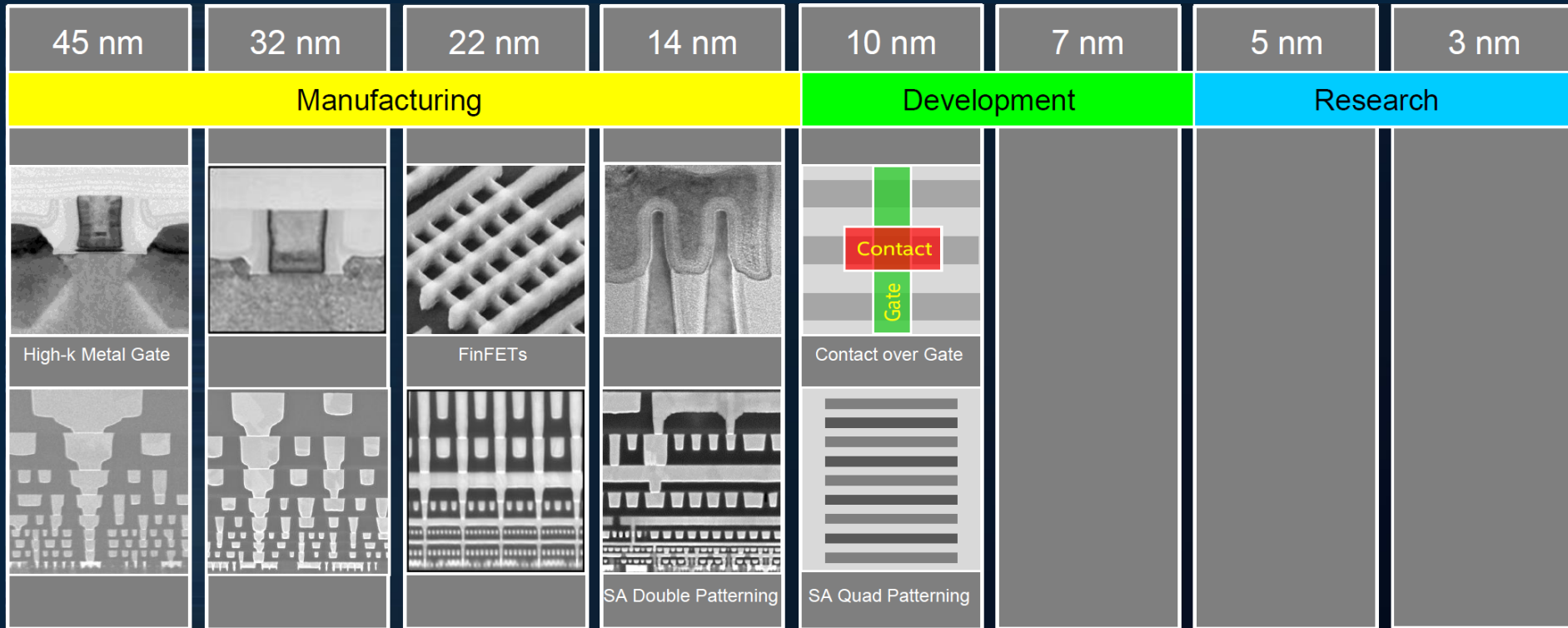
What is needed:

- Narrow fins
- Mobility enhancement
- Conductive metals
- Tunable materials
- Precision structure control

Source: Applied Materials, Semicon West, 2013

<https://www.extremetech.com/computing/162376-7nm-5nm-3nm-the-new-materials-and-transistors-that-will-take-us-to-the-limits-of-moores-law>

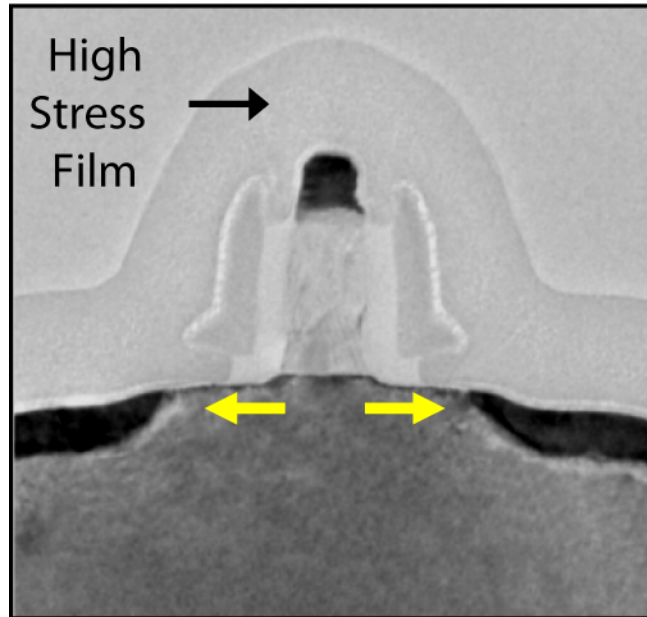
Innovation Enabled Technology Pipeline



- The development of each new technology node has required true technological innovations

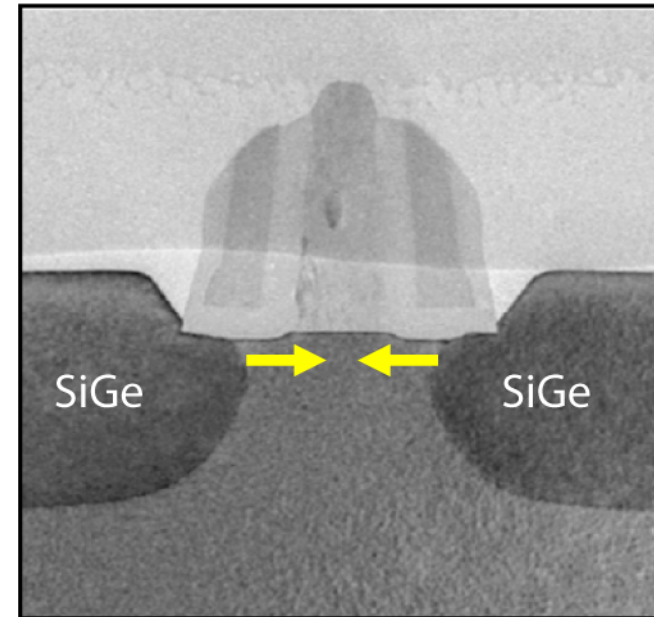
90 nm Strained Silicon Transistors

NMOS



SiN cap layer
Tensile channel strain

PMOS

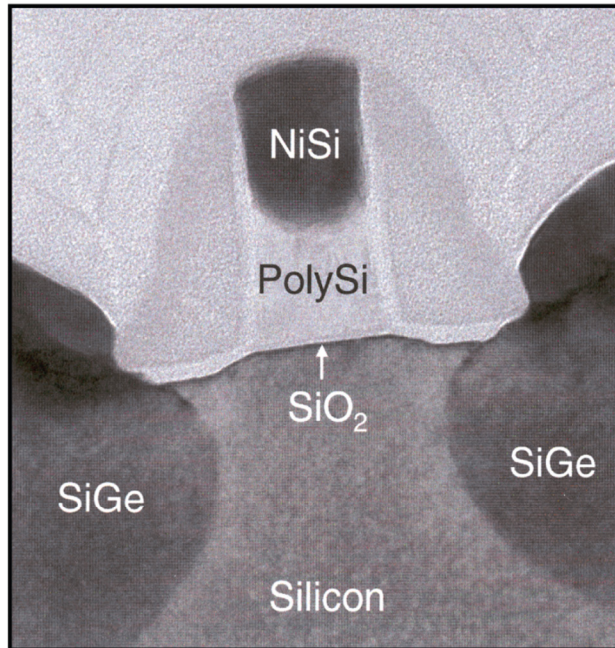


SiGe source-drain
Compressive channel strain

- Strained silicon provided increased drive currents by boosting the mobility, making up for lack of gate oxide scaling

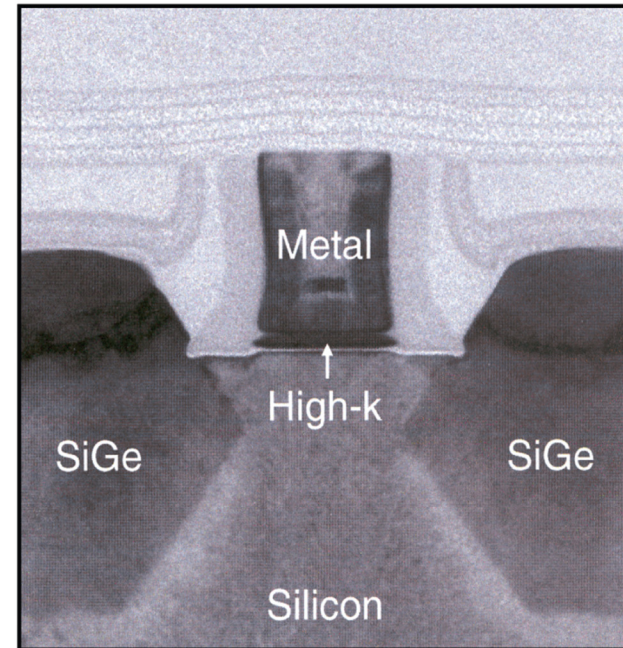
High-k + Metal Gate Transistors

65 nm Transistor



SiO₂ dielectric
Polysilicon gate electrode

45 nm HK+MG



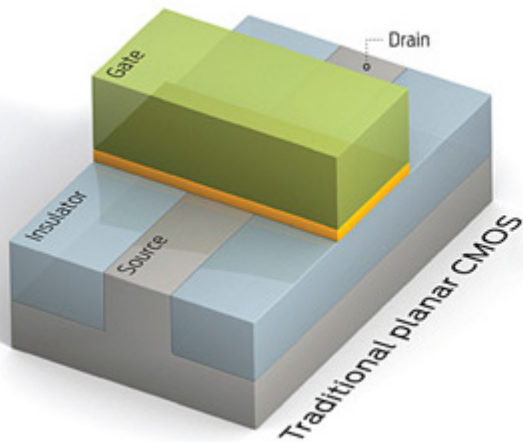
Hafnium-based dielectric
Metal gate electrode

- 65 nm technology node
- 1.2 nm gate oxide
- 35 nm effective gate length

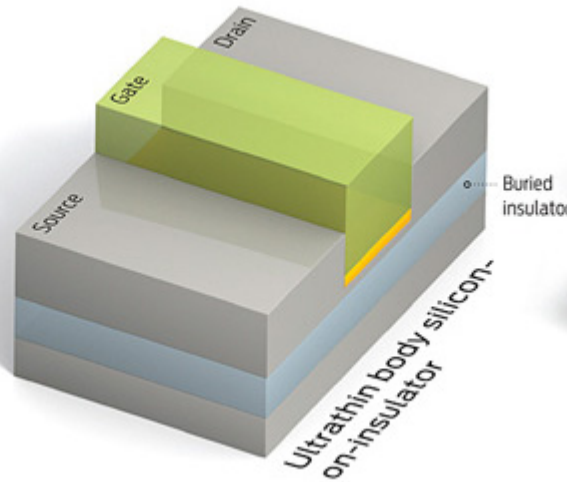
- High-k + metal gate transistors break through gate oxide scaling barrier

Transistor War – FinFET versus UTB-SOI

Bulk MOSFET



UTB-SOI Ultrathin Body Silicon-on-Insulator



FinFET

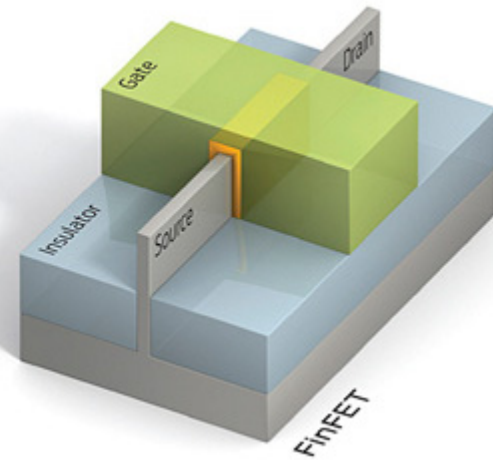
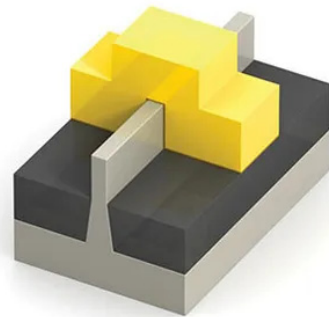
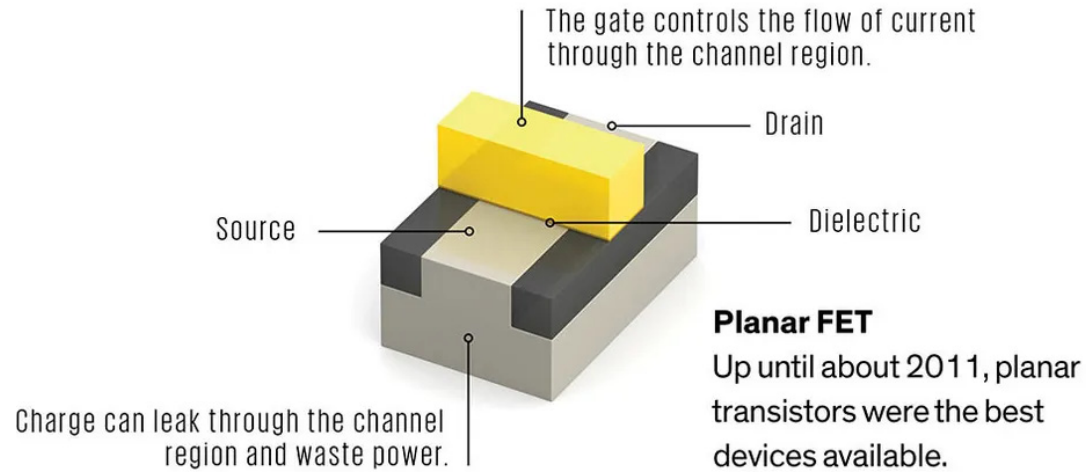
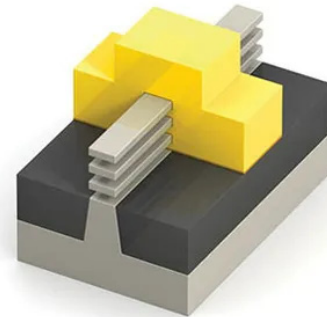


Illustration: Emily Cooper

Transistor Architecture Evolution

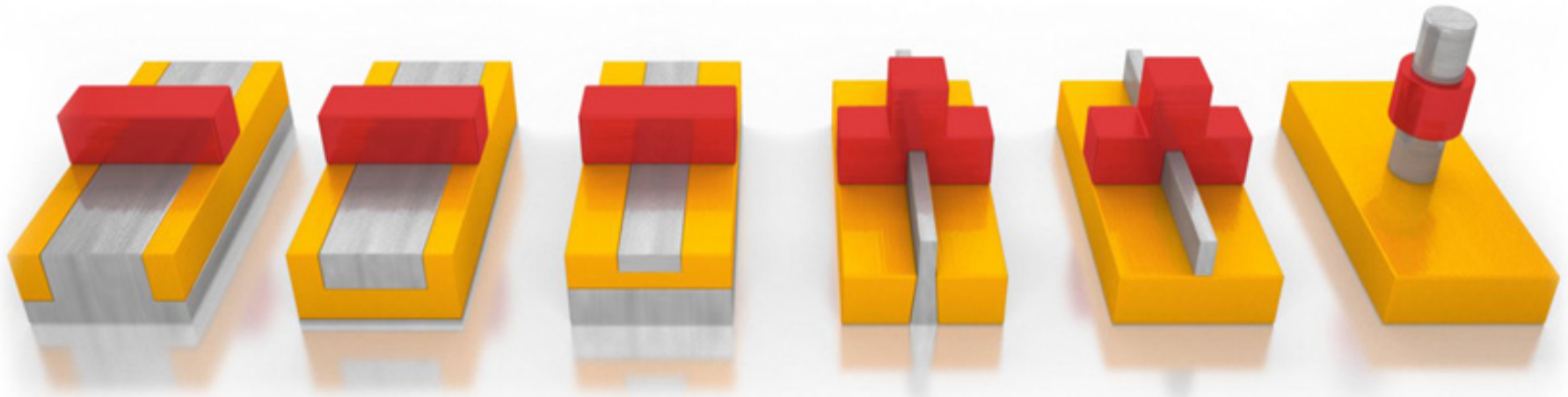


FinFET
Surrounding the channel region on three sides with the gate gives better control and prevents current leakage.



Stacked nanosheet FET
The gate completely surrounds the channel regions to give even better control than the FinFET.

No End in Sight for Logic Scaling



N 20

N 20 / N 14

N10

N 20 / N 7

N 7 / N 5

N 5 / N 3.5

Bulk CMOS:
Complementary
Metal Oxide
Semiconductor

SOI: Partially
depleted Silicon
on insulator

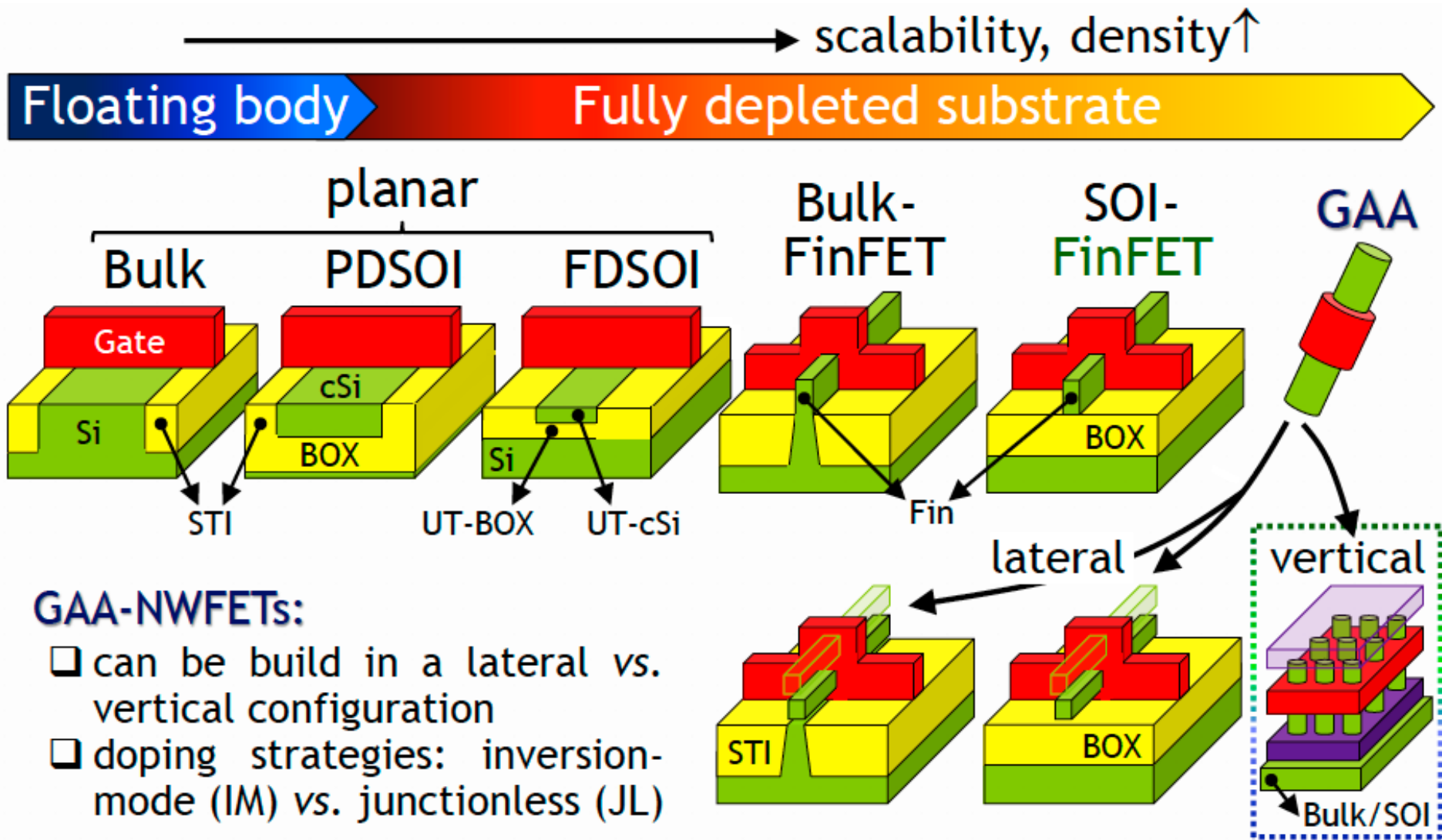
SOI: Fully depleted
Silicon on insulator

Bulk FinFet :
fin field effect
transistor

SOI FinFet :
silicon on insulator
fin field effect
transistor, III-V

Gate-all-around
transistor

Scaling Scenario for Device Architectures (1/2)



Source: International Roadmap for Devices and Systems (IDRS), 2022 Edition, Executive Summary.

Scaling Scenario for Device Architectures (2/2)

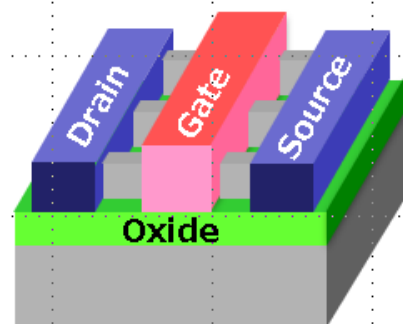
>2020: 2.5D/3D fine-pitch assembly + stacking →

FinFET
2011-2022

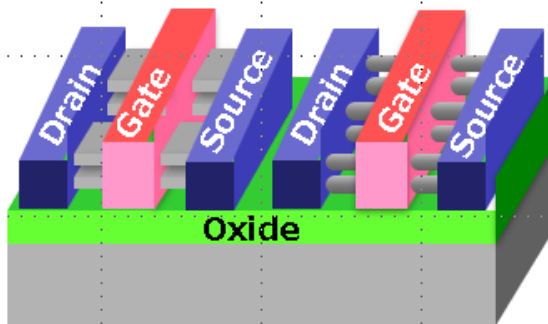
Lateral GAA
2022-2037

CFET
2028-2037

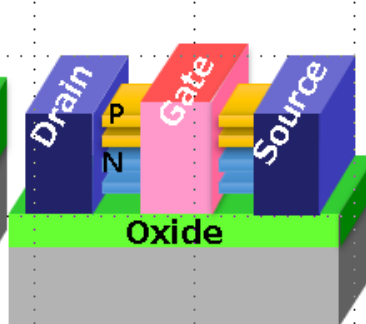
3D VLSI
2031-2037



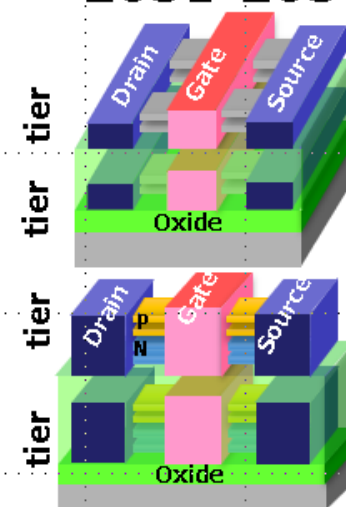
- Increasing drive by taller fin
- Better channel control for better perf-power



- Increasing drive by stacked devices
- Better channel control
- Reduced footprint stdcell



- Increased stacking
- Reduced PN proximity
- Reduced footprint stdcell



- Sequential heterogenous integration/fine-pitch stacking (e.g., logic, memory, NVM, analog, IO, RF, sensors)

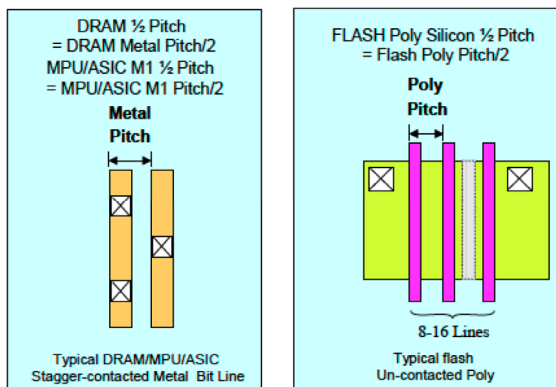
imec's Potential Roadmap Extension



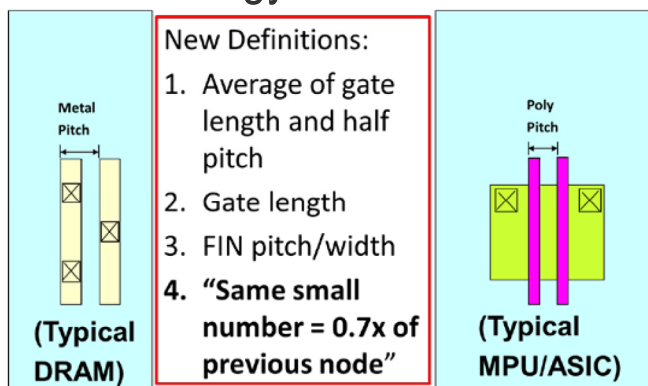
Source: International Roadmap for Devices and Systems (IDRS), 2022 Edition, Executive Summary.

Evolving Node Definitions

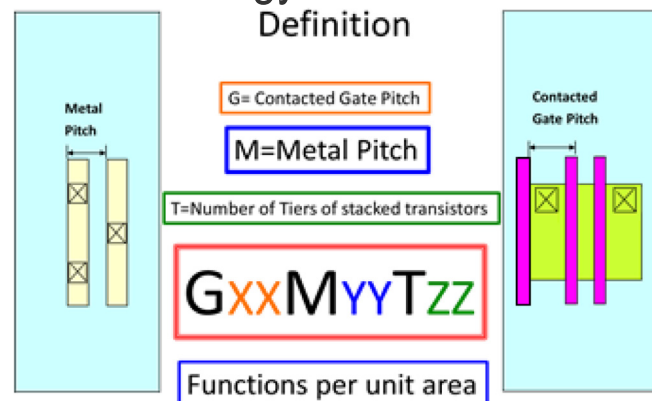
Original technology node definition



Industry “adaptation” of technology node definition



IRDS comprehensive technology node definition

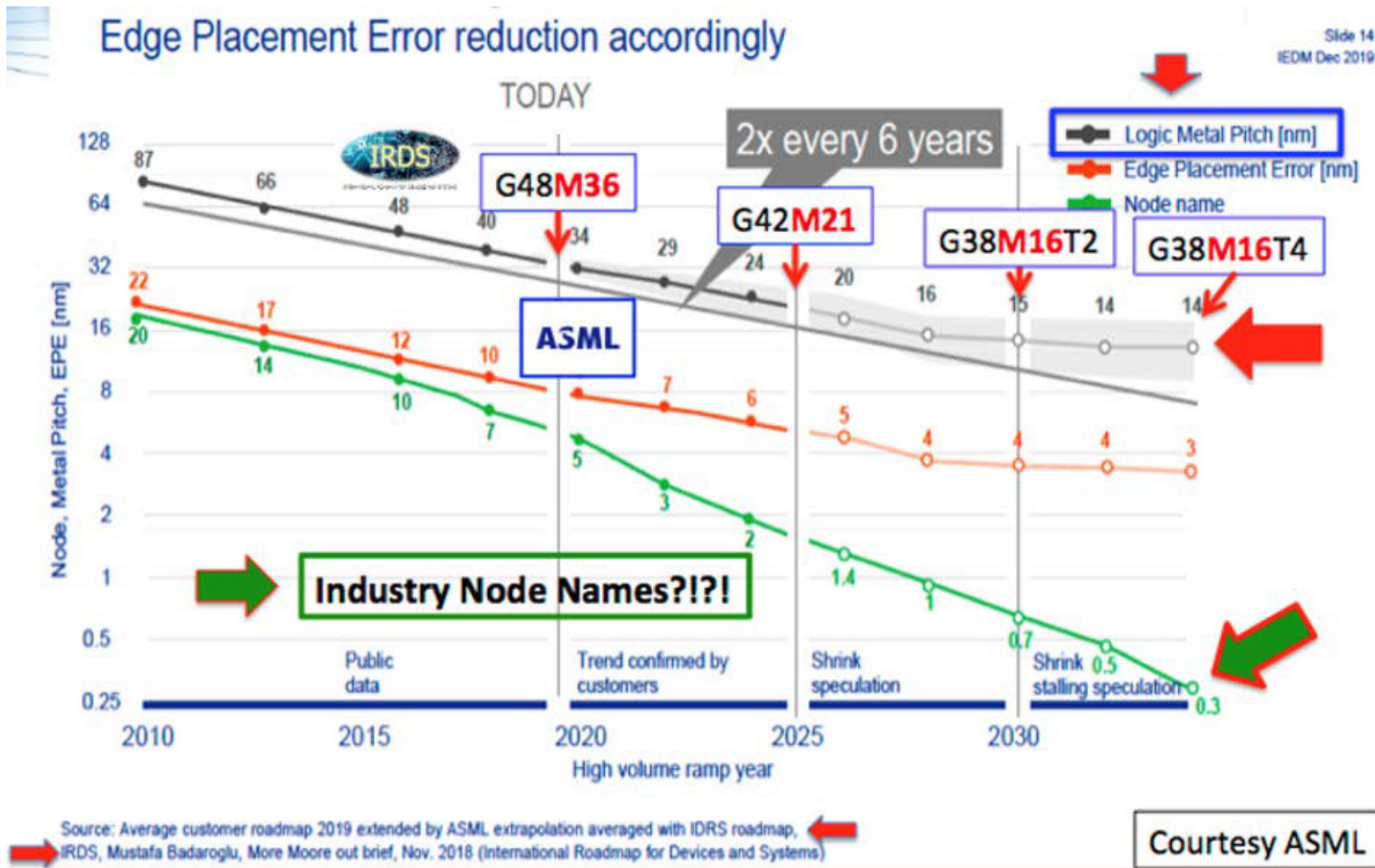


Device Scaling in Coming Years

| YEAR OF PRODUCTION | 2022 | 2025 | 2028 | 2031 | 2034 | 2037 |
|---|----------------|----------|-------------------|-------------------------|-------------------------|-------------------------|
| Logic industry "Node Range" Labeling | G48M24 | G45M20 | G42M16 | G40M16 T2 | G38M16 T4 | G38M16 T6 |
| Fine-pitch 3D integration scheme | Stacking | Stacking | Stacking | 3DVLSI | 3DVLSI | 3DVLSI |
| Logic device structure options | finFET LGAA | LGAA | LGAA CFET-SRAM | LGAA-3D CFET-SRAM | LGAA-3D CFET-SRAM | LGAA-3D CFET-SRAM |
| Platform device for logic | finFET | LGAA | LGAA CFET-SRAM | LGAA-3D CFET-SRAM-3D | LGAA-3D CFET-SRAM-3D | LGAA-3D CFET-SRAM-3D |
| | | | | | | |
| LOGIC DEVICE GROUND RULES | | | | | | |
| Mx pitch (nm) | 32 | 24 | 20 | 16 | 16 | 16 |
| M1 pitch (nm) | 32 | 23 | 21 | 20 | 19 | 19 |
| M0 pitch (nm) | 24 | 20 | 16 | 16 | 16 | 16 |
| Gate pitch (nm) | 48 | 45 | 42 | 40 | 38 | 38 |
| Lg: Gate Length - HP (nm) | 16 | 14 | 12 | 12 | 12 | 12 |
| Lg: Gate Length - HD (nm) | 18 | 14 | 12 | 12 | 12 | 12 |
| Channel overlap ratio - two-sided | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 |
| Spacer width (nm) | 6 | 6 | 5 | 5 | 4 | 4 |
| Spacer k value | 3.5 | 3.3 | 3.0 | 3.0 | 2.7 | 2.7 |
| Contact CD (nm) - finFET, LGAA | 20 | 19 | 20 | 18 | 18 | 18 |
| Device architecture key ground rules | | | | | | |
| Device lateral pitch (nm) | 24 | 26 | 24 | 24 | 23 | 23 |
| Device height (nm) | 48 | 52 | 48 | 64 | 60 | 56 |
| FinFET Fin width (nm) | 5.0 | | | | | |
| Footprint drive efficiency - finFET | 4.21 | | | | | |
| Lateral GAA vertical pitch (nm) | | 18.0 | 16.0 | 16.0 | 15.0 | 14.0 |
| Lateral GAA (nanosheet) thickness (nm) | | 6.0 | 6.0 | 6.0 | 5.0 | 4.0 |
| Number of vertically stacked nanosheets on one device | | 3 | 3 | 4 | 4 | 4 |
| LGAA width (nm) - HP | | 30 | 30 | 20 | 15 | 15 |
| LGAA width (nm) - HD | | 15 | 10 | 10 | 6 | 6 |
| LGAA width (nm) - SRAM | | 7 | 6 | 6 | 6 | 6 |
| Footprint drive efficiency - lateral GAA - HP | | 4.41 | 4.50 | 5.47 | 5.00 | 4.75 |
| Device effective width (nm) - HP | 101.0 | 216.0 | 216.0 | 208.0 | 160.0 | 152.0 |
| Device effective width (nm) - HD | 101.0 | 126.0 | 96.0 | 128.0 | 88.0 | 80.0 |
| PN seperation width (nm) | 45 | 40 | 20 | 15 | 15 | 10 |

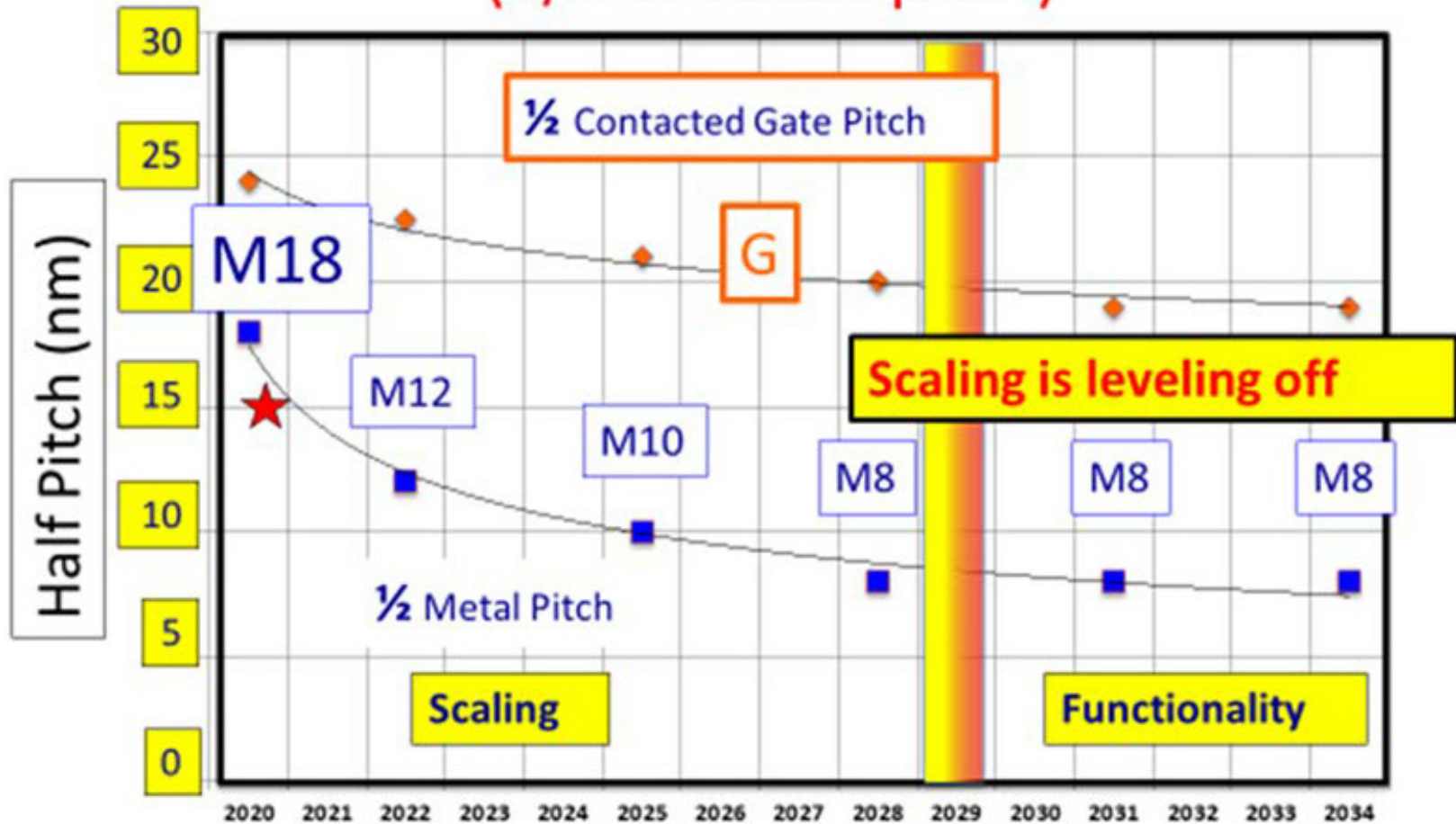
Source: International Roadmap for Devices and Systems (IDRS), 2022 Edition, Executive Summary.

Dimensional Scaling Continues another Decade



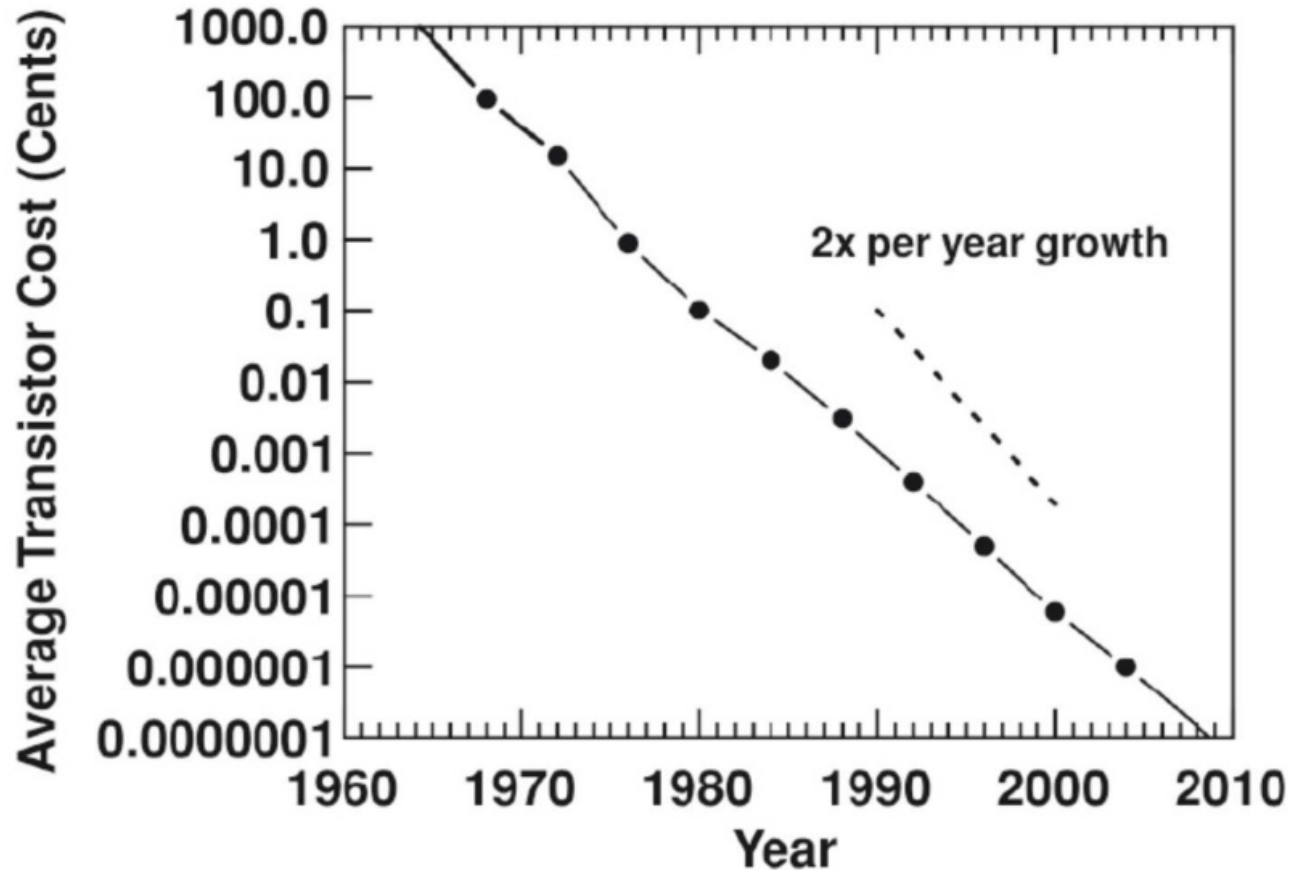
Scaling will Reach Fundamental Limits around 7-8 nm

Reconnecting with NTRS/ITRS Technology Nodes
(1/2 of metal pitch)



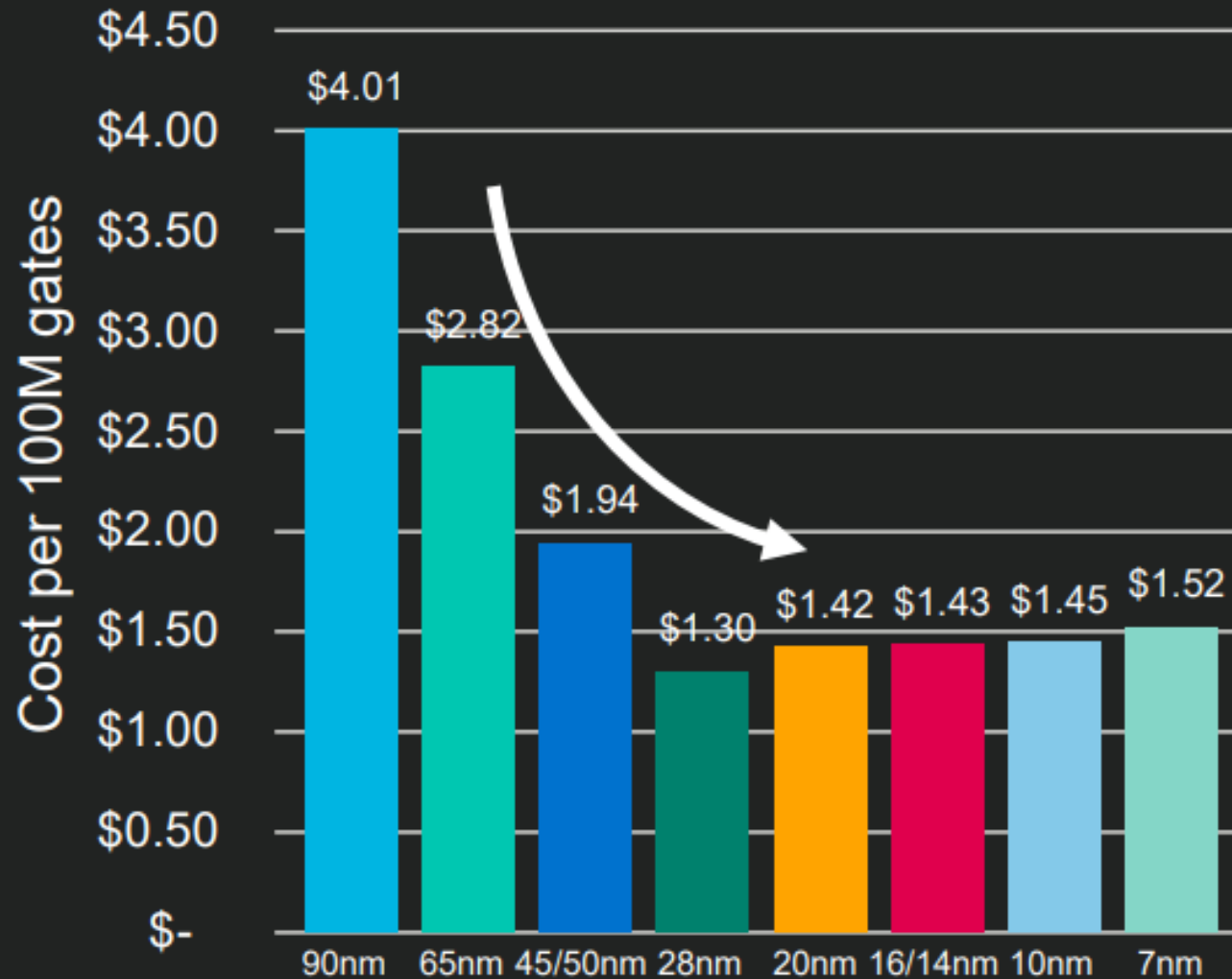
Source: International Roadmap for Devices and Systems (IDRS), 2022 Edition, Executive Summary.

It's All About Economics...



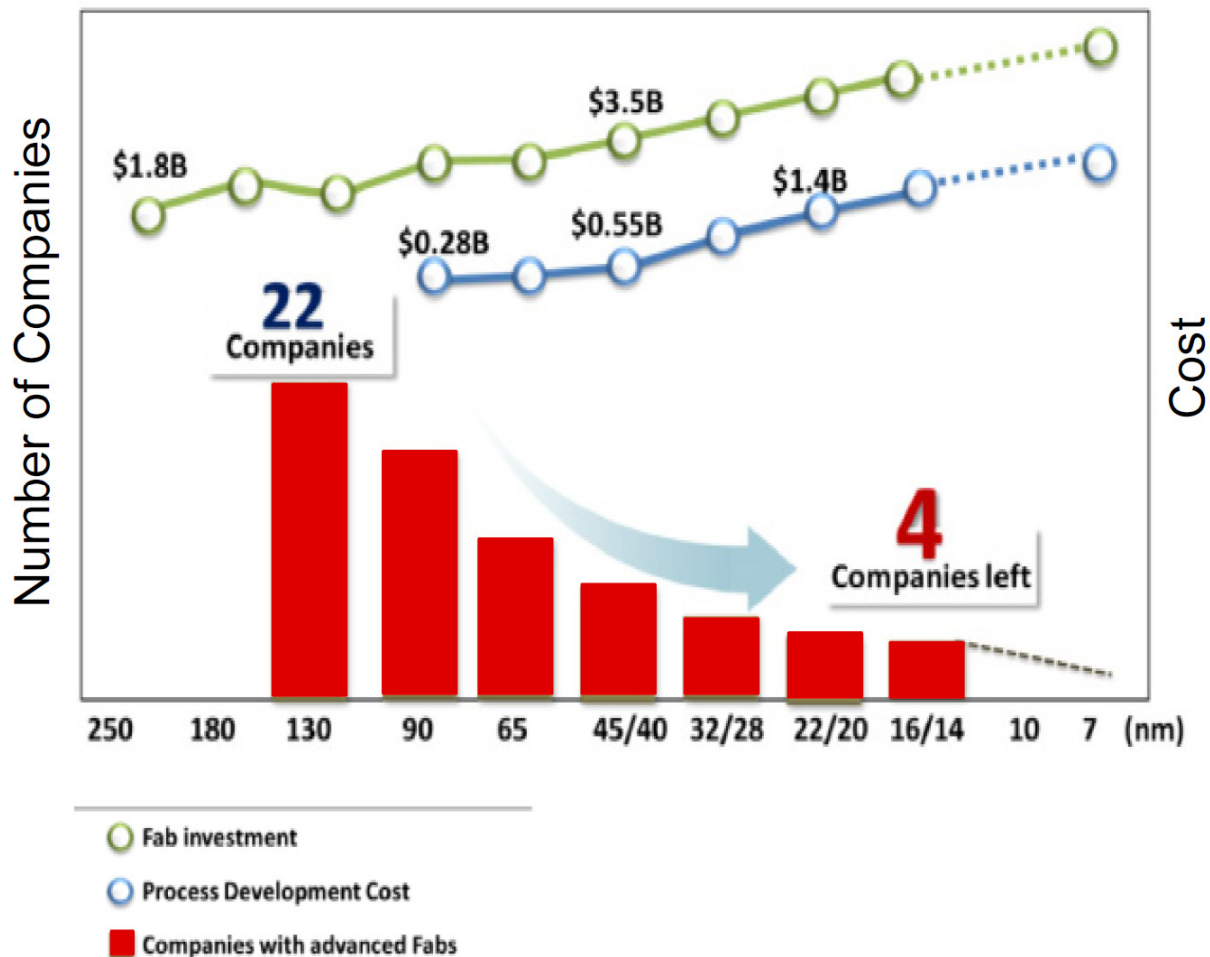
- The average price of a single transistor has fallen almost by 10^{10} in less than 5 decades!

Cost per Transistor is Rising



Source: <https://www.fabricatedknowledge.com/p/the-rising-tide-of-semiconductor>.

Smaller Number of Players for Leading Edge Nodes



EE Times Connecting the Global Electronics Community
designlines SoC

News & Analysis

GlobalFoundries Halts 7nm Work

Next FinFET node would have cost \$2-4B

Rick Merritt

8/27/2018 04:00 PM EDT

17 comments

NO RATINGS
LOGIN TO RATE

Like 124 Tweet in Share G+

SAN JOSE, Calif. – The race to drive semiconductor technology to the bleeding edge has narrowed to three companies.

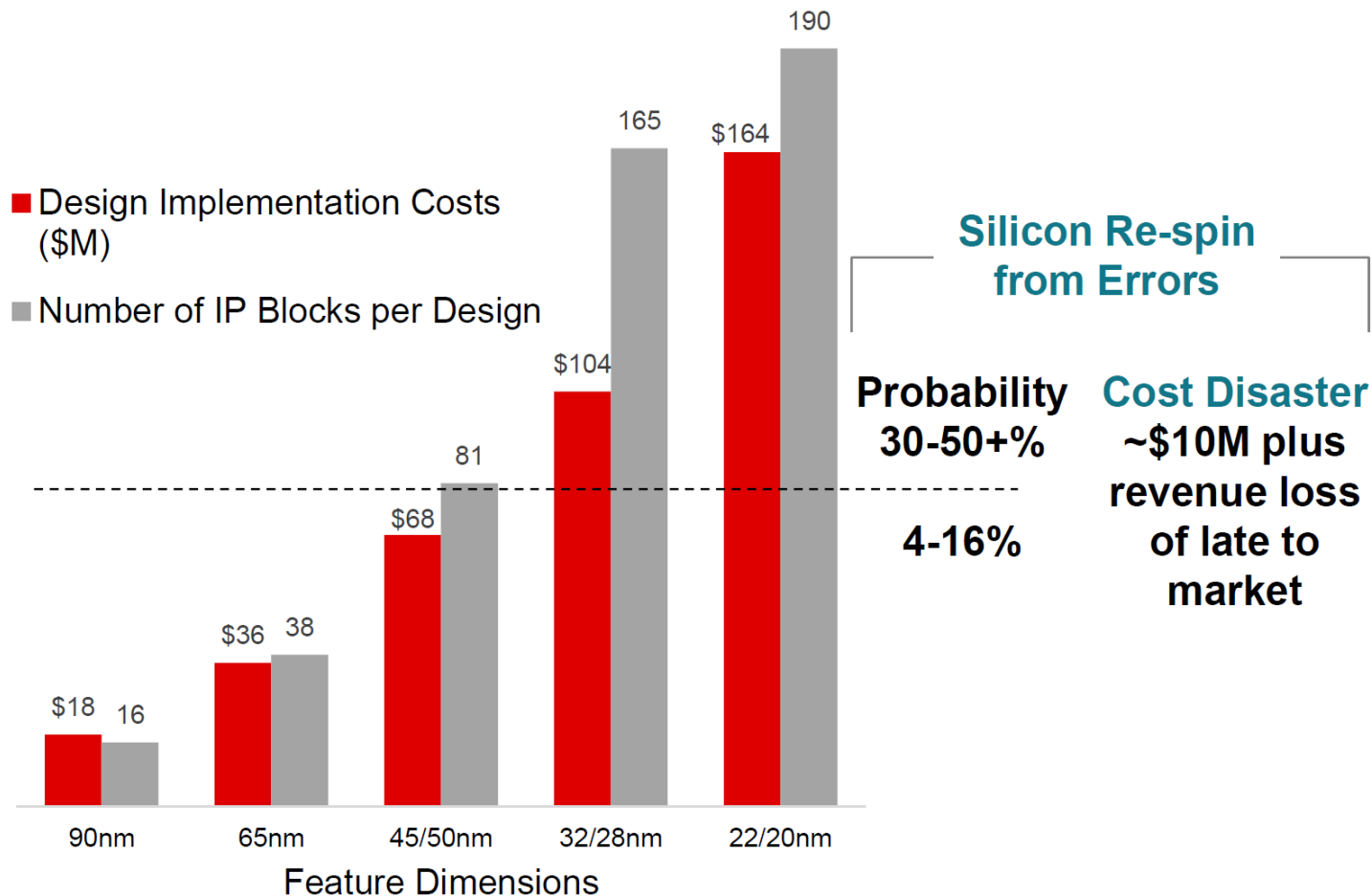
Globalfoundries suspended work on a 7nm node. It will lay off less than 5% of its workforce and make its ASIC group a wholly-owned subsidiary so it can partner with one of the remaining 7nm foundries.

It would have cost GF \$2-4 billion to ramp up the 40-50,000 wafers/month capacity needed to have a chance of making a return on the node. "The financial investment didn't make as much sense as doing something else," said Tom Caulfield, the former general manager of Fab 8 named chief executive of GF in March.

Source: Samsung Foundry data.

A. Bahai, ISSCC 2017.

Rising Design Cost and Complexity



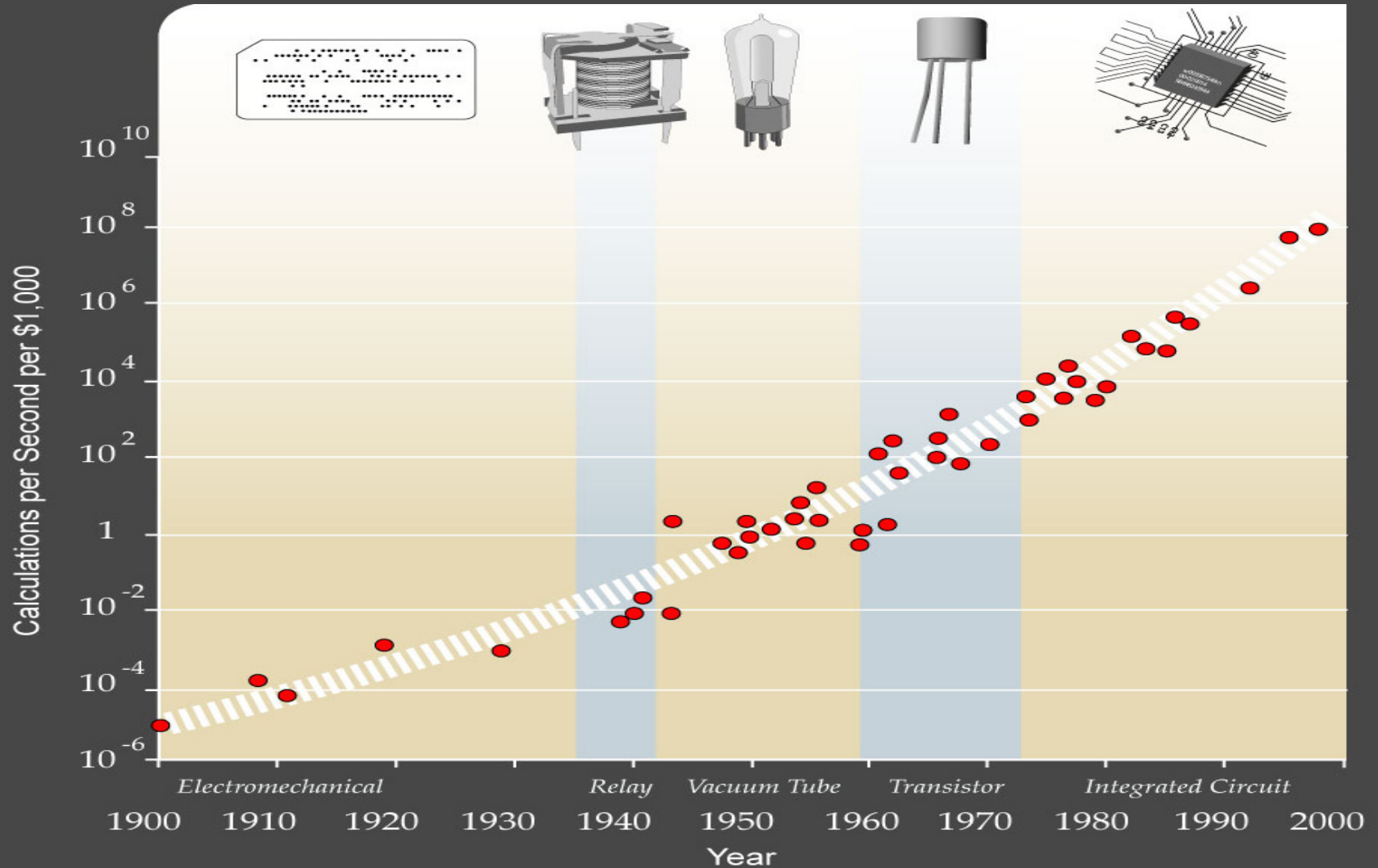
Source: International Business Strategies.

A. Bahai, ISSCC 2017.

Outline

- Introduction
- CMOS technology scaling
- **Power consumption and energy efficiency**
- Voltage scaling

Computation Power Evolution



Computation Efficiency – The Koomey Law

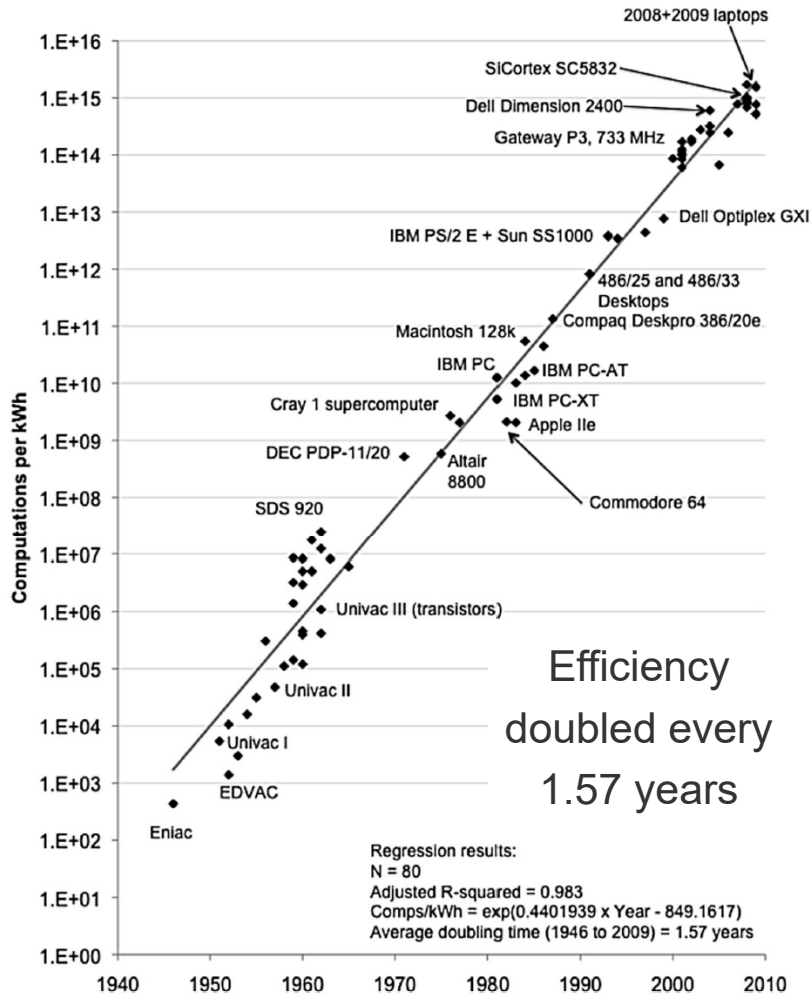


Figure 3. Computations per kilowatt-hour over time. These data include a range of computers, from PCs to mainframe computers and measure computing efficiency at peak performance. Efficiency doubled every 1.57 years from 1946 to 2009.

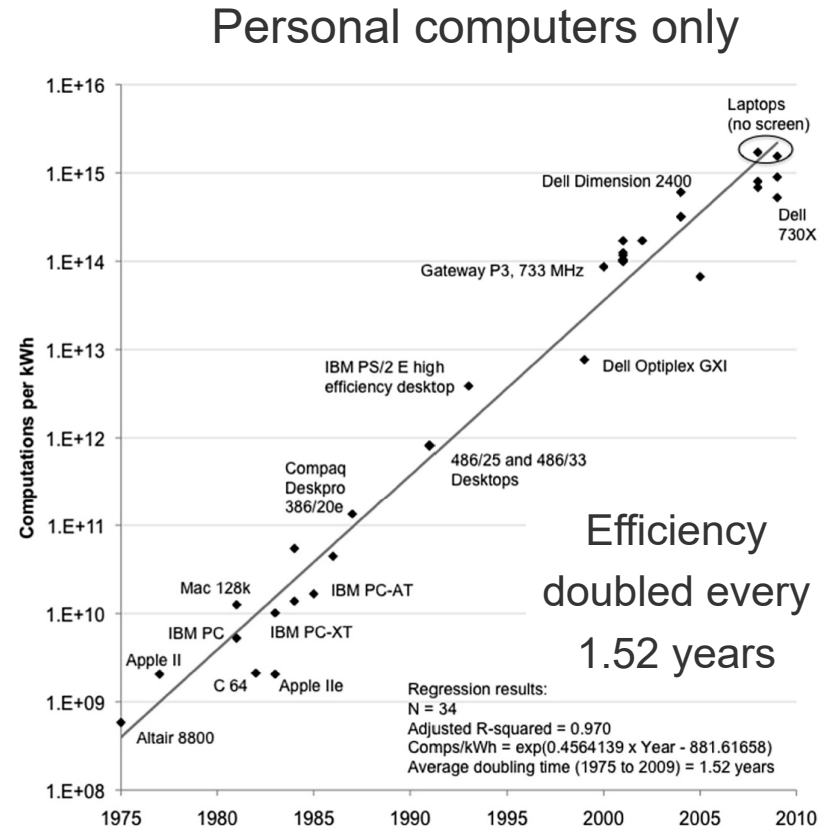


Figure 4. Computations per kilowatt-hour over time for personal computers alone. Efficiency doubled every 1.52 years from 1975 to 2009.

Extending the Koomey Law by 3D Integration

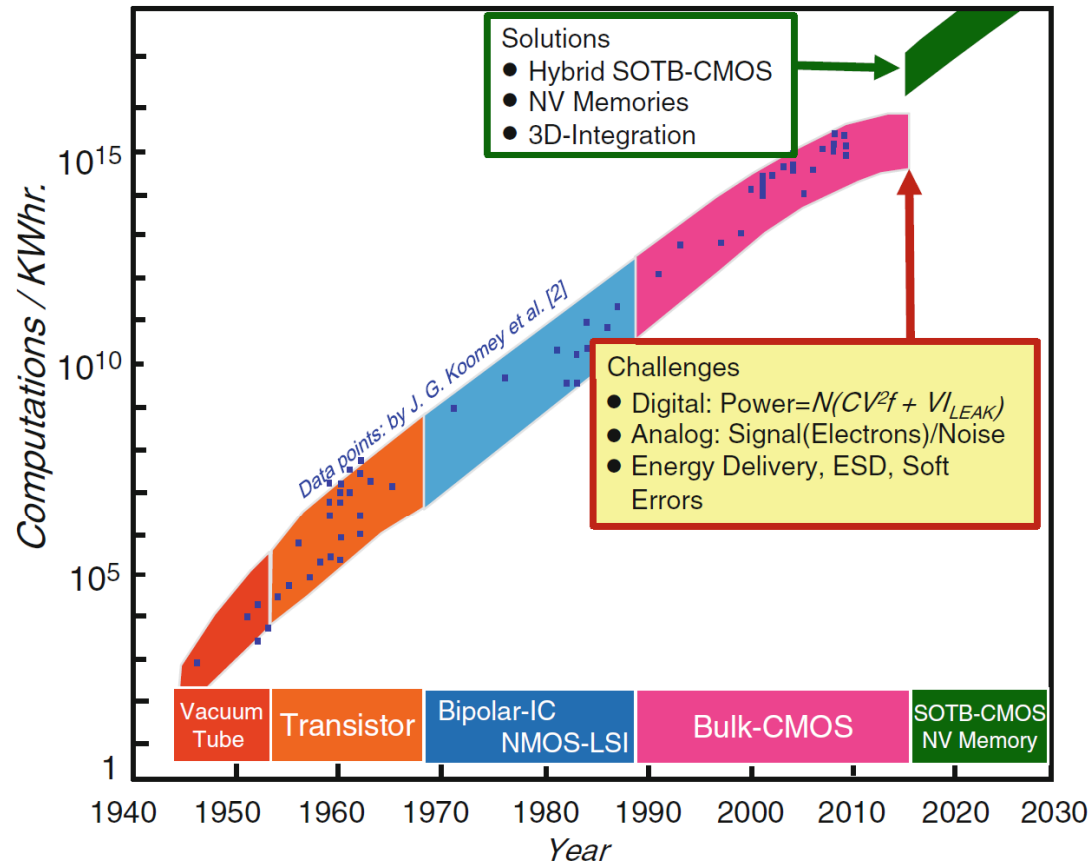
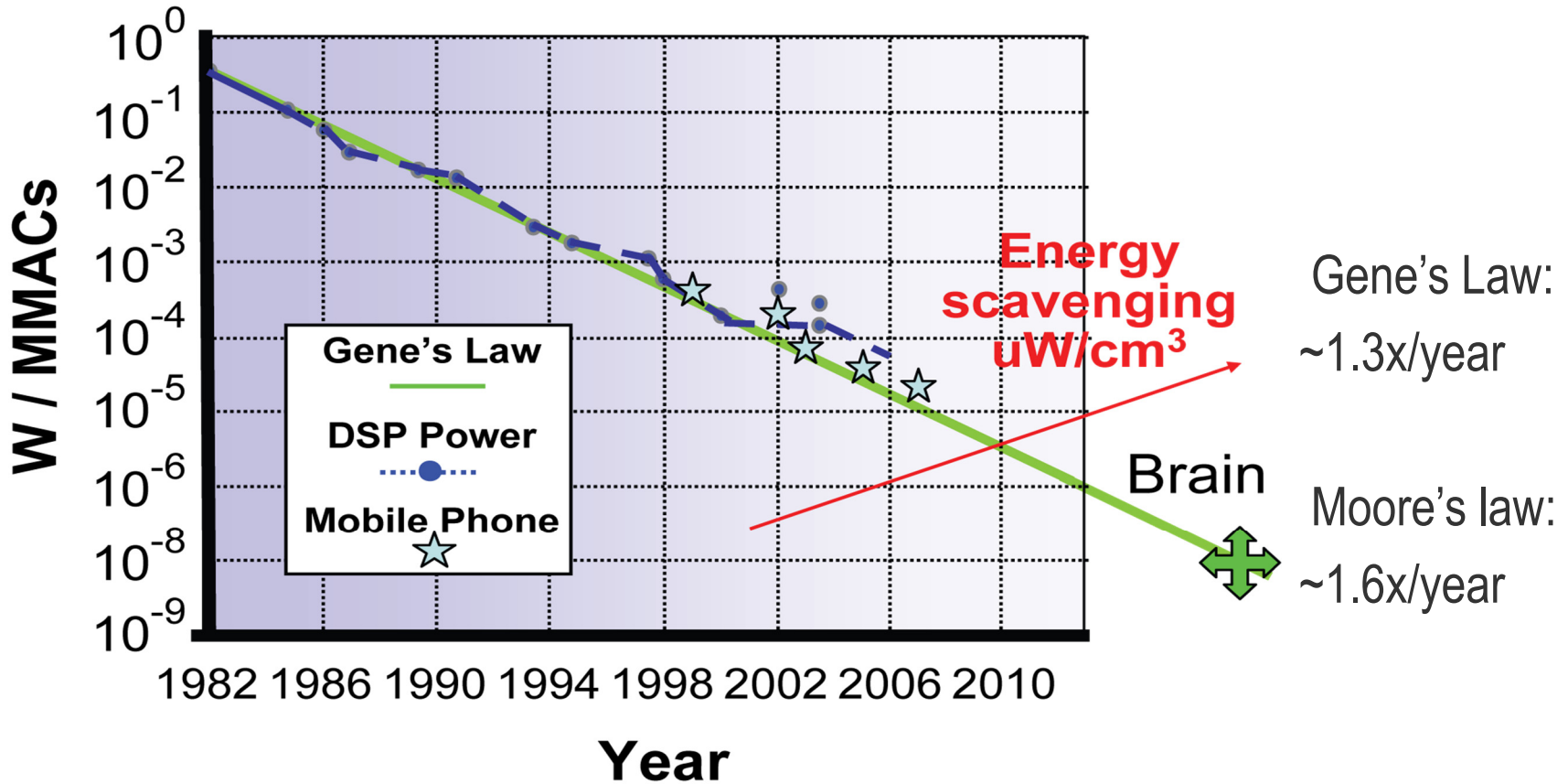


Fig. 2.2 Evolution of the computations/kWh and the major device technologies. Data points are taken from Koomey's paper [2]. Bulk CMOS needs to be exchanged by another technology solution by 2020 due to the power-dissipation limit (© 2009 IEEE)

Gene's Law for DSPs



- Power dissipation per MAC operation has decreased by half every 18 months
- Does not hold for analog processing and data communication

Source: Gene A. Frantz, TI Developer Conf., 2008.

G. Frantz, "Digital signal processor trends," IEEE Micro, vol. 20, no. 6, pp. 52-59, Nov.-Dec. 2000.

Switching Energy

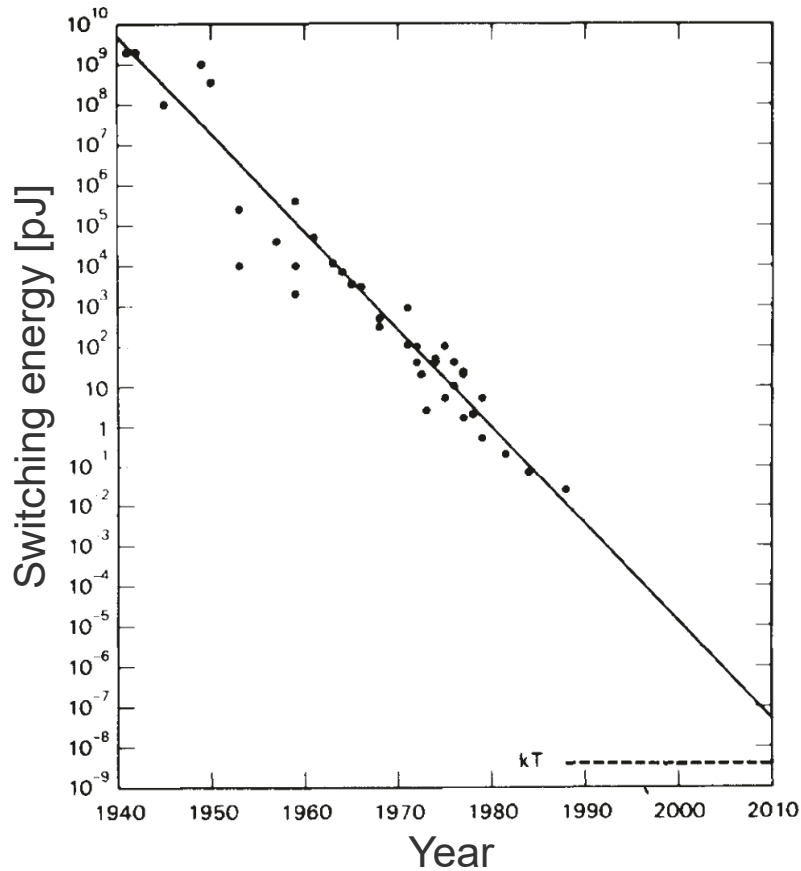
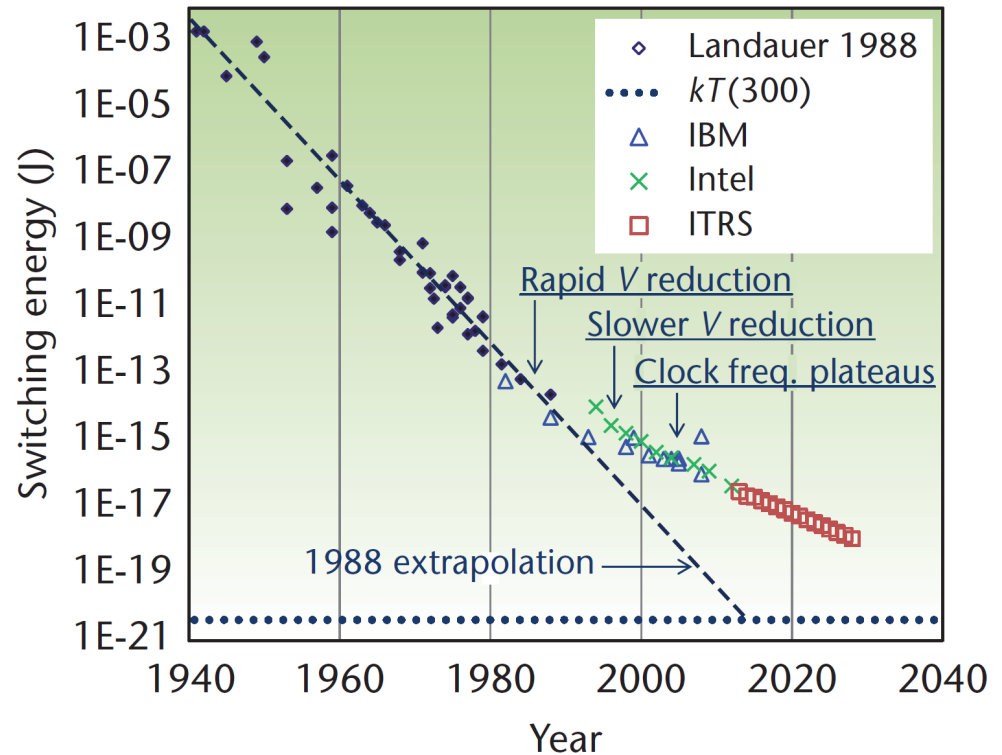


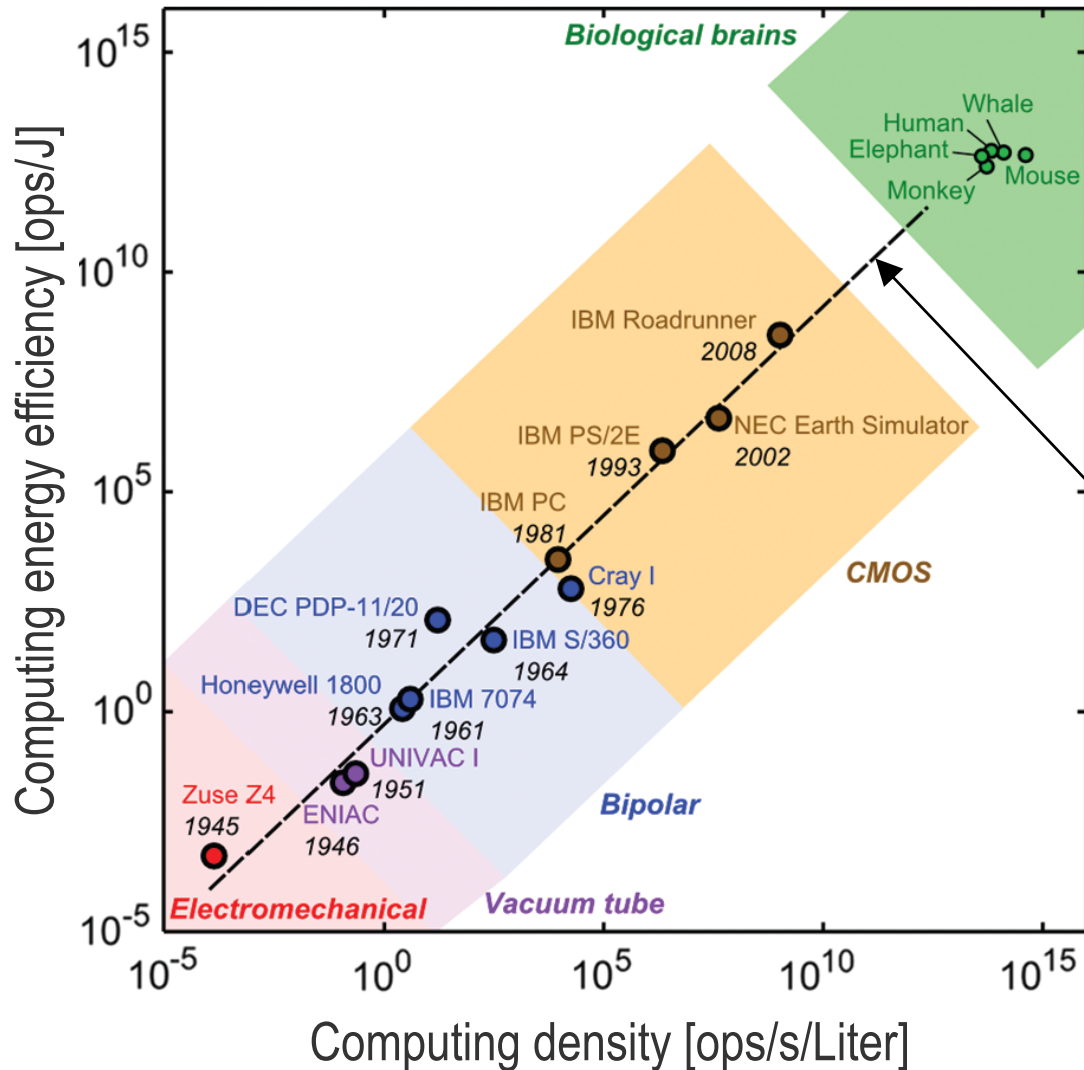
Fig. 1 The decrease in energy dissipated per logic operation over recent decades.

📖 R. Landauer, Nature, vol. 335, no. 6193, pp. 779-784, Oct. 1988.



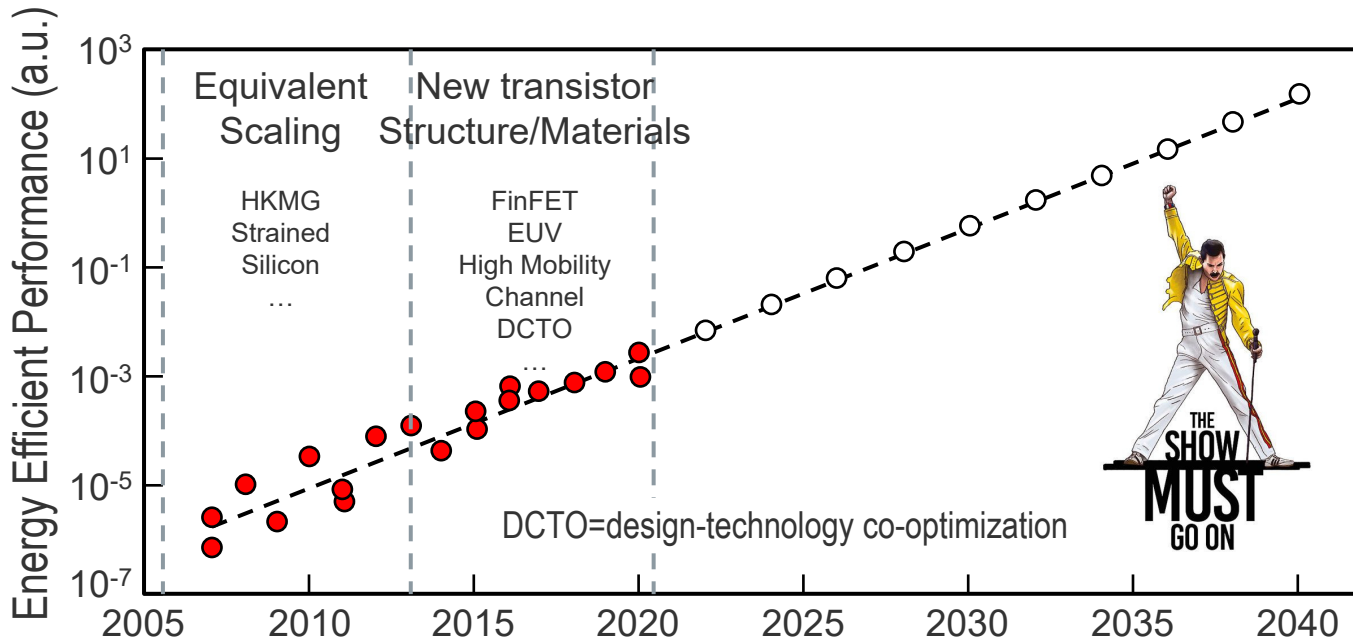
📖 T. N. Theis and H. P. Wong, Computing in Science & Engineering, vol. 19, no. 2, pp. 41-50, May/June 2017.

Computing Efficiency versus Computing Density



Lack of scalability of VLSI architectures in the 3rd dimension

The show must go on...

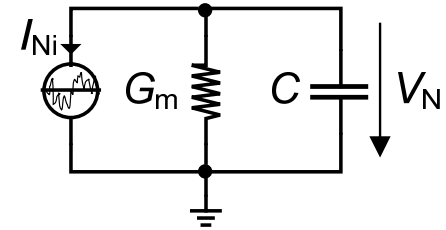
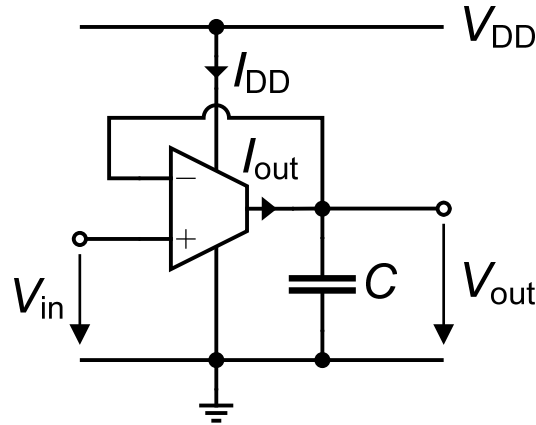
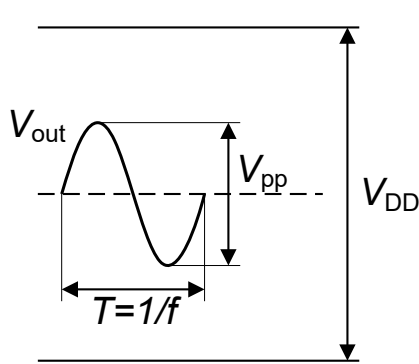


Technology node **N45 N28 N16 N10 N7 N5 N3** Year

$$\text{Energy-efficient performance} = \text{Throughput} \times \text{Throughput/Watt} [1/(\text{fJ.ps})]$$

- Historical trend (<2020) and projection of energy efficient performance gains showing an expected $\sim 2\times$ improvement every 2 years

Lower Limit of Power Consumption (1/2)

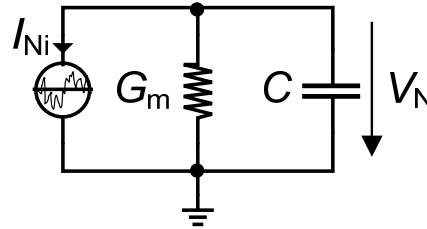


Output noise current model with
 $S_{Ni} = 4kT \cdot \gamma \cdot G_m$

- Assumptions: 100% efficient transconductor (i.e. $I_{out} = I_{DD}$)
- Low-voltage usually comes at the cost of a **higher power consumption**
- What are the fundamental lower **limits** to power consumption?
- Average value of I_{out} is given by $\overline{I_{out}} = f \cdot C \cdot V_{pp}$
- The average power consumption P is then given by

$$P = V_{DD} \cdot f \cdot C \cdot V_{pp} = \frac{V_{DD}}{V_{pp}} \cdot f \cdot C \cdot V_{pp}^2$$

Lower Limit of Power Consumption (2/2)



- The **noise current** power spectral density (PSD) is given by

$$S_{Ni} = 4kT \cdot \gamma \cdot G_m$$

- The total mean square **noise voltage** across capacitor C is given by

$$V_N^2 = \frac{\gamma \cdot kT}{C}$$

- Where γ is the **noise excess factor** which will be assumed to be unity
- The **signal-to-noise** ratio SNR is then given by

$$SNR = \frac{V_{pp}^2/8}{kT/C} \Rightarrow V_{pp}^2 = \frac{8kT}{C} \cdot SNR$$

- The **power consumption** can then be written as

$$P = 8 \frac{V_{DD}}{V_{pp}} \cdot kT \cdot f \cdot SNR$$

Limit of Power Consumption – Factor of Merit

- Power consumption is minimized by maximizing the peak-to-peak signal with **rail-to-rail operation** $V_{pp} = V_{DD}$

$$P_{min} = 8kT \cdot f \cdot SNR$$

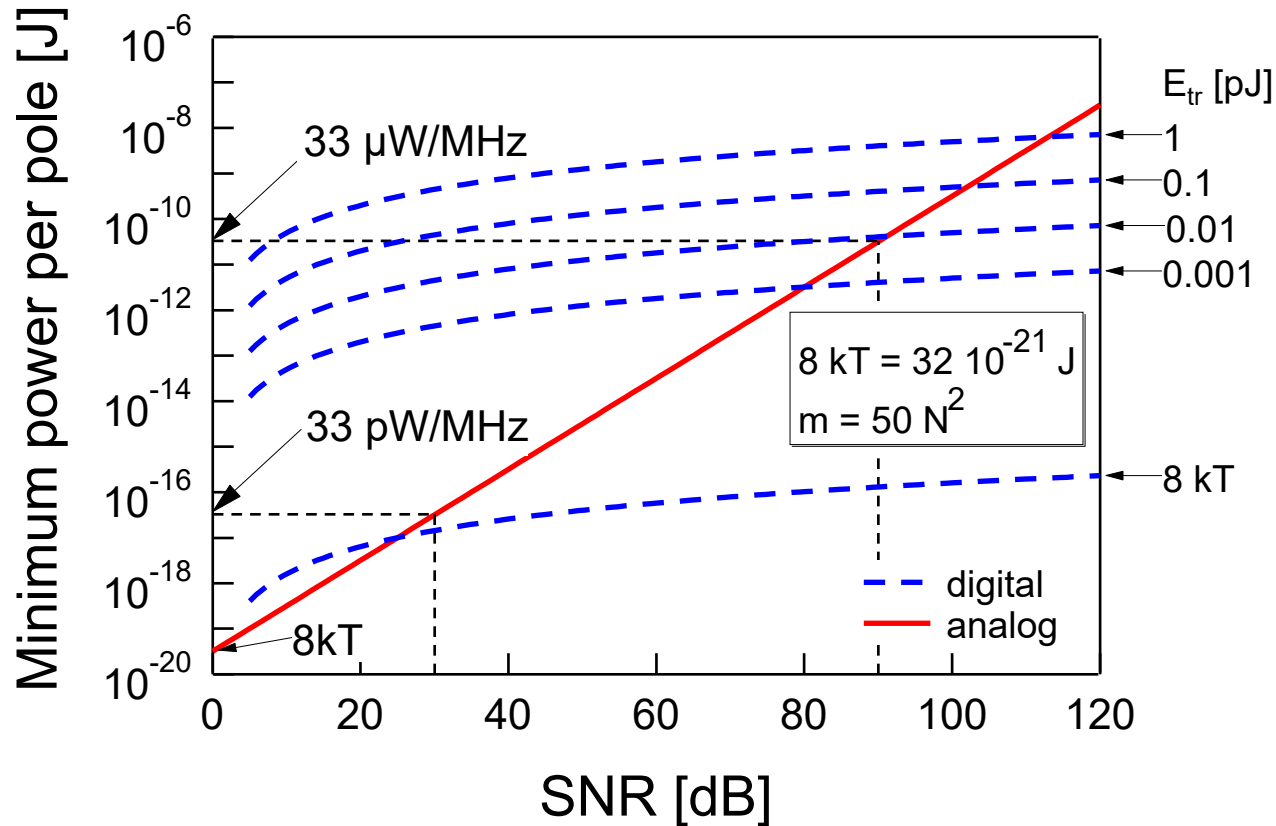
- P is proportional to frequency which actually corresponds to the **bandwidth B** for low-pass filters
- A factor of merit (actually demerit, the smaller the better) can be defined as

$$K \triangleq \frac{P}{kT \cdot B \cdot SNR} = 8 \frac{V_{DD}}{V_{pp}}$$

- K is **minimum** for $V_{pp} = V_{DD}$ (rail-to-rail linear operation)

$$K_{min} = K \Big|_{V_{pp}=V_{DD}} = 8$$

Minimum Power Consumption versus SNR



- The **minimum** power consumption P_{min} is proportional to frequency (bandwidth)

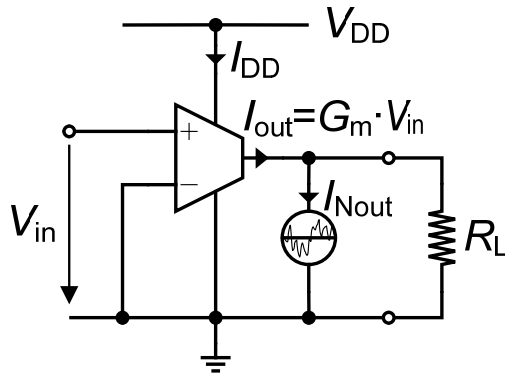
$$P_{min} = 8kT \cdot f \cdot SNR$$

- It corresponds to an **absolute minimum** for processing a signal with an analog circuit

Practical Power Limitations

- K_{min} constitute an **absolute minimum** not accounting for many non-idealities
- In practical analog circuits there are many **non-idealities** that can seriously degrade (increase) the factor K far beyond K_{min}
 - ▶ **Current inefficiency** (non-ideal class B operation)
 - ▶ **Linearity** requirement
 - ▶ Additional **bias** circuits
 - ▶ Limited **matching**
 - ▶ **Additional noise** contributions (from flicker noise and from other devices)
 - ▶ Parasitic **capacitances**
 - ▶ Charge injection

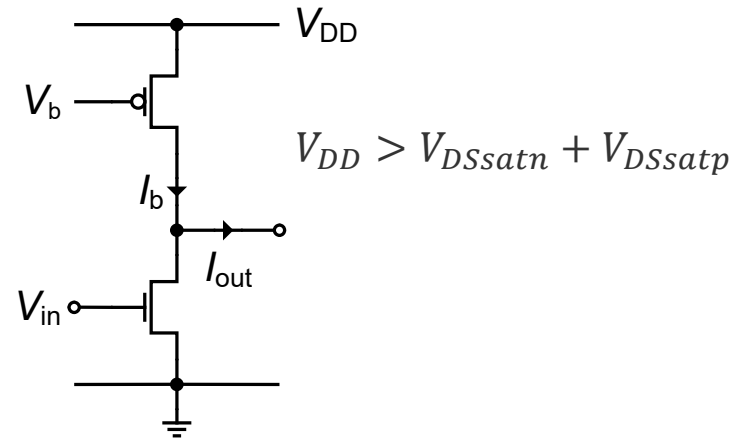
Minimum Power of a Transconductance Amplifier



$$S_{I_{out}^2} = 4kT \cdot \gamma_n \cdot G_m$$

$$V_N^2 = \frac{\gamma_n kT}{C}$$

$$B = \frac{G_m}{C}$$



- K factor of a **generic transconductor** is given by

$$K \triangleq \frac{P}{kT \cdot B \cdot SNR} \geq 4\gamma_n \frac{V_{DD}}{V_{in,rms}^2} \frac{I_{DD}}{G_m}$$

- Can be minimized by maximizing V_{in}/V_{DD} (**rail-to-rail operation**) and G_m/I_{DD} (**bias in weak inversion**)

- Simple NMOS transconductor biased in SI for better linearity
- K minimum for $V_{DSsatn} = V_{DSsatp} = V_{DSSat}$

$$K > 8\gamma_n n \left(\frac{V_{DSSat}}{V_{in,rms}} \right)^2$$

- Can be minimized by reducing V_{DSSat} and hence the supply voltage V_{DD}

Min. Power Cons. for NMOS Transconductor

- However, decreasing V_{DSSat} increases the **total harmonic distortion** THD due to the square-law characteristic according to

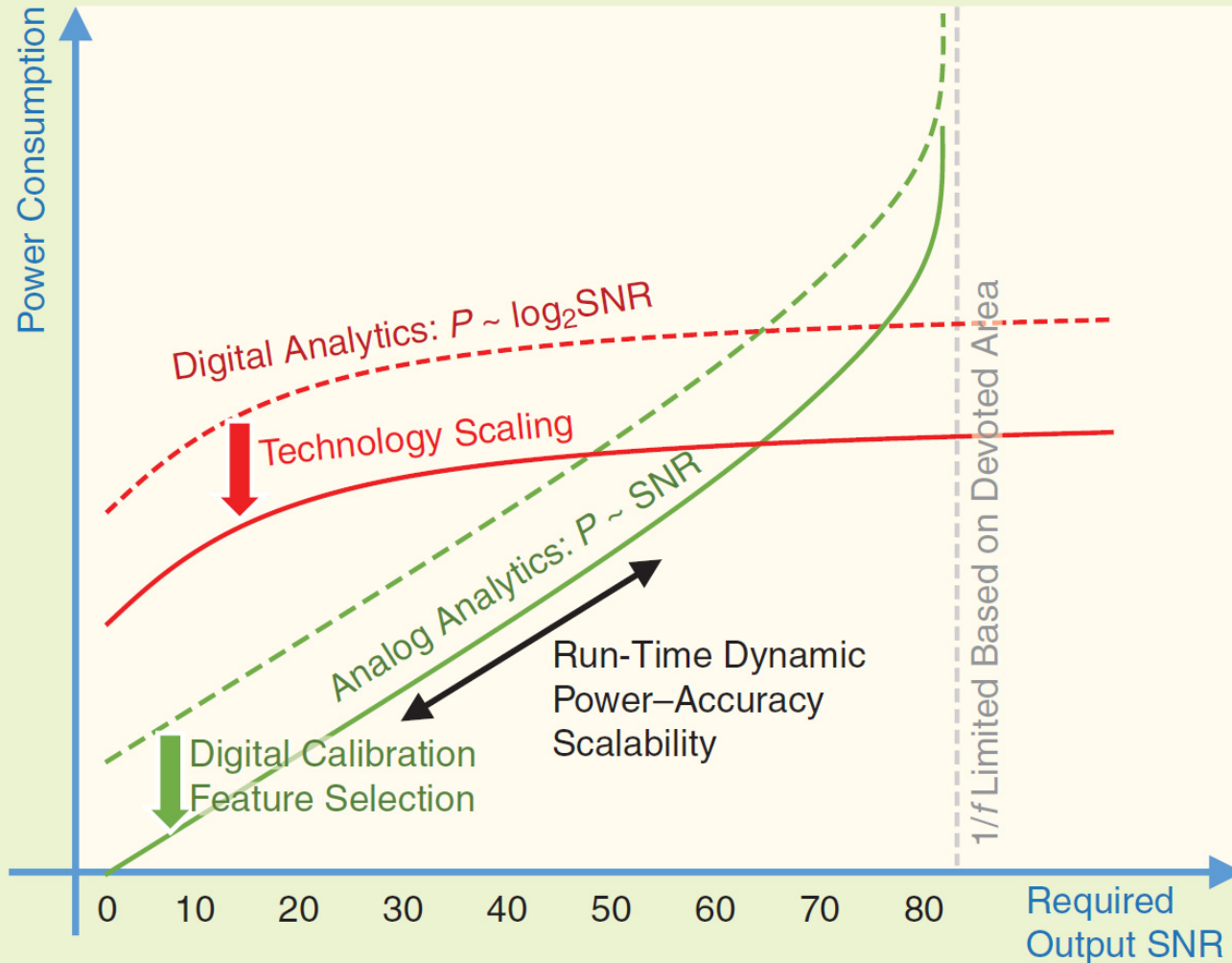
$$THD = \frac{V_{in}}{4nV_{DSSat}} = \frac{\sqrt{2}V_{in,rms}}{4nV_{DSSat}} \Rightarrow \left(\frac{V_{DSSat}}{V_{in,rms}} \right)^2 = \frac{1}{8n^2THD^2}$$

- K can then be expressed directly in terms of the THD as

$$K > \frac{\gamma}{THD^2} = \frac{2}{3THD^2}$$

- Having $THD < 1\%$ results in $K > 6700$ instead of 8 (**factor 840 higher!**)

Analog and Digital Power Consumption

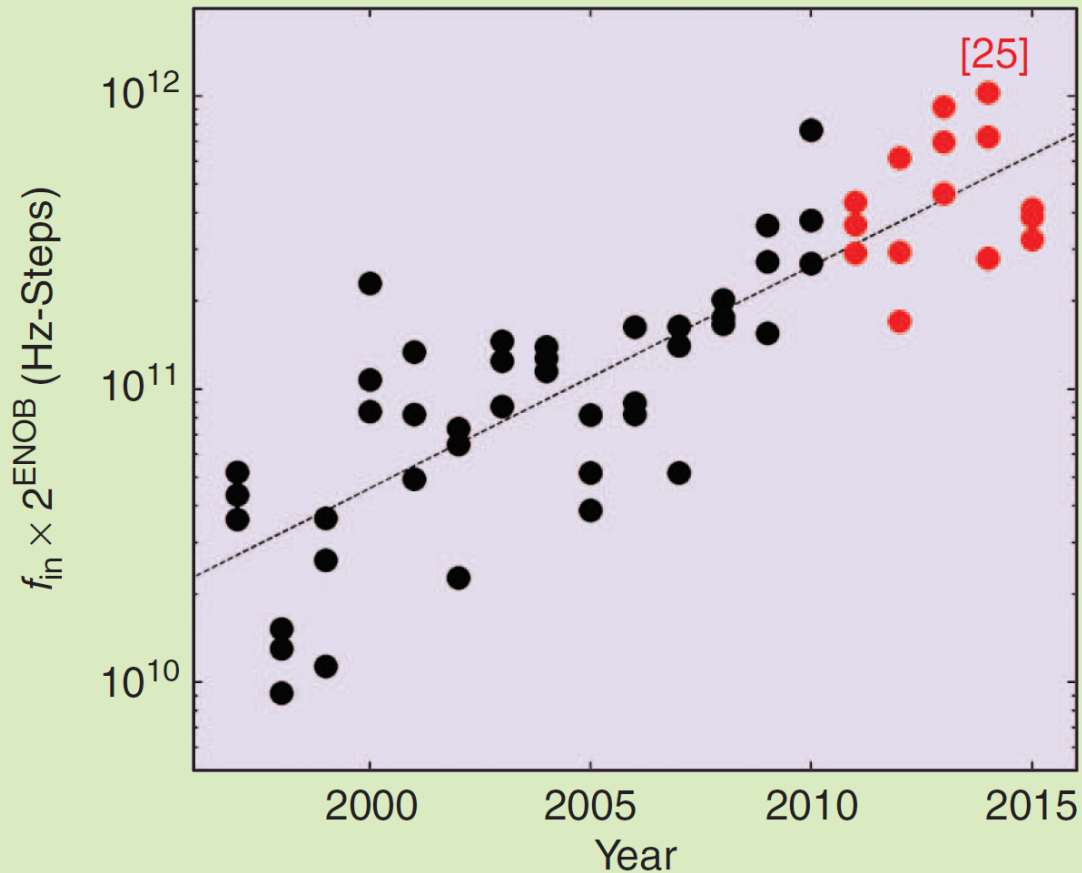


E. A. Vittoz, "Future of analog in the VLSI environment," ISCAS 1990.

C. Enz and E. Vittoz, CMOS Low-Power Analog Circuit Design in *Emerging technologies: Designing Low Power Digital Systems*, Wiley 1996.

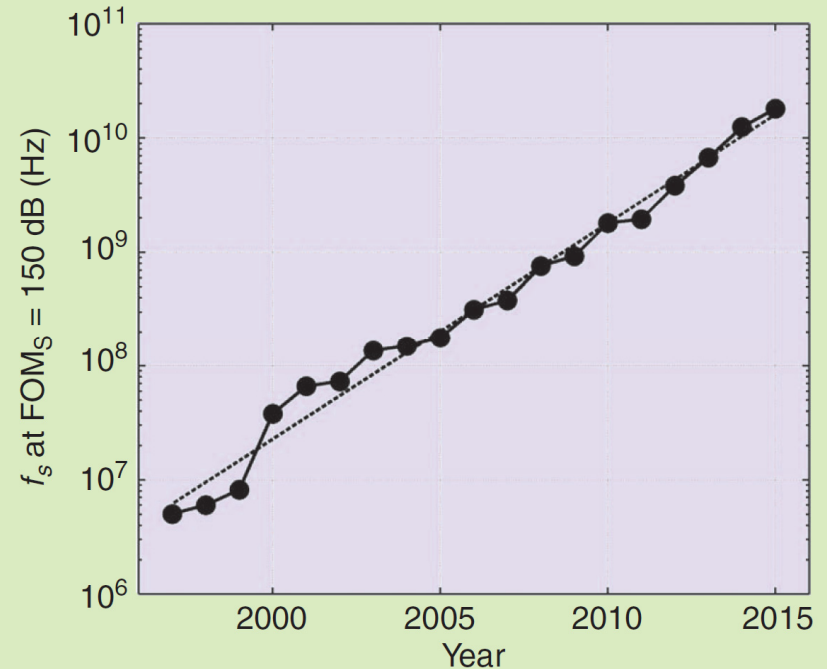
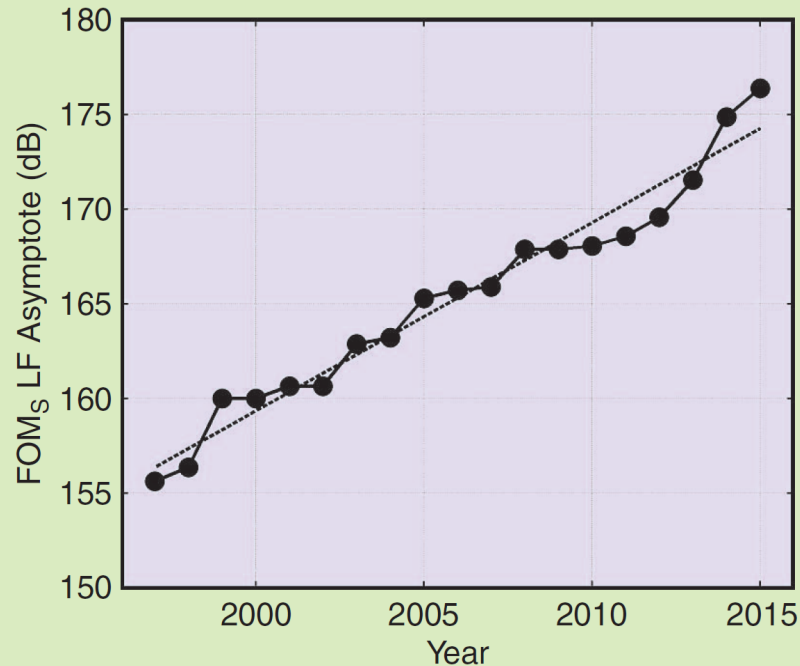
M. Verhelst and A. Bahai, "Where Analog Meets Digital," IEEE Solid-State Circuits Society Magazine, Summer 2015.

Evolution of ADCs Speed-resolution Product



- The speed-resolution product has doubled every four years
- Note that f_{in} represents the maximum input frequency of the signal to be converted

ADCs FoMs Trends



- Schreier FoM defined as

$$FoM_S = SNDR(dB) + 10 \log \left(\frac{f_s/2}{P} \right)$$

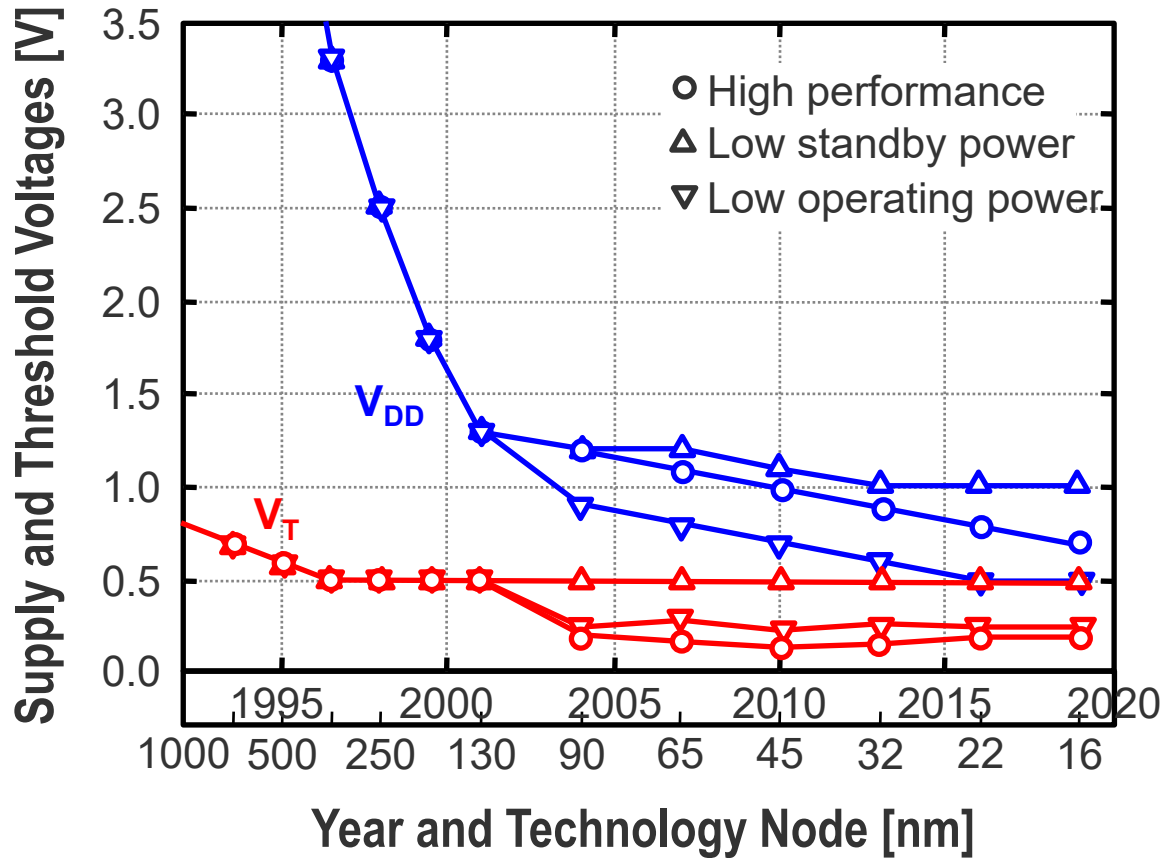
- Progression at 1dB/year or doubling the energy efficiency every three years

- Conversion rate is doubling every 1.8 years or 60x improvement every ten years

Outline

- Introduction
- CMOS technology scaling
- Power consumption and energy efficiency
- **Voltage scaling**

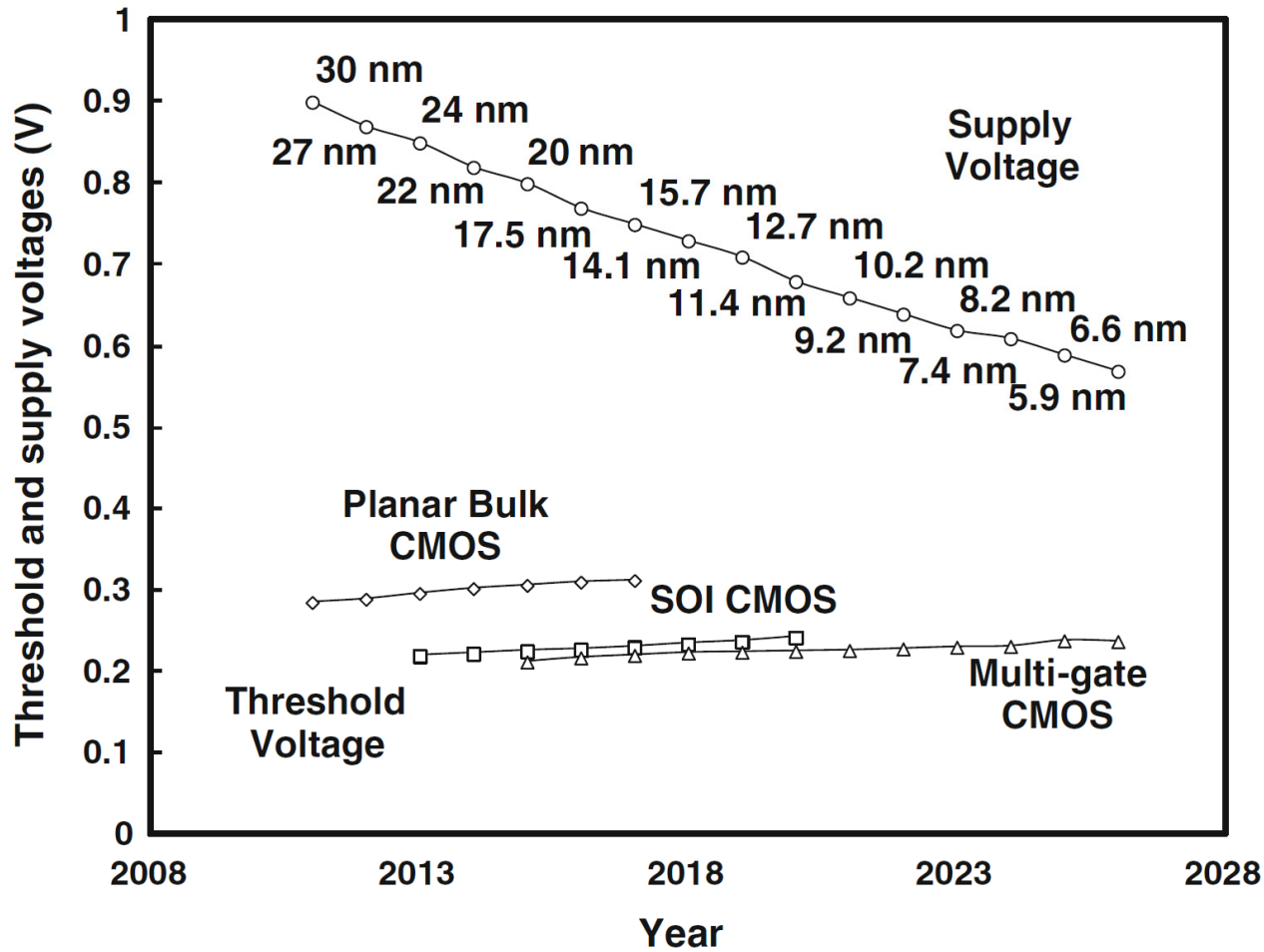
The ITRS Roadmap (2005)



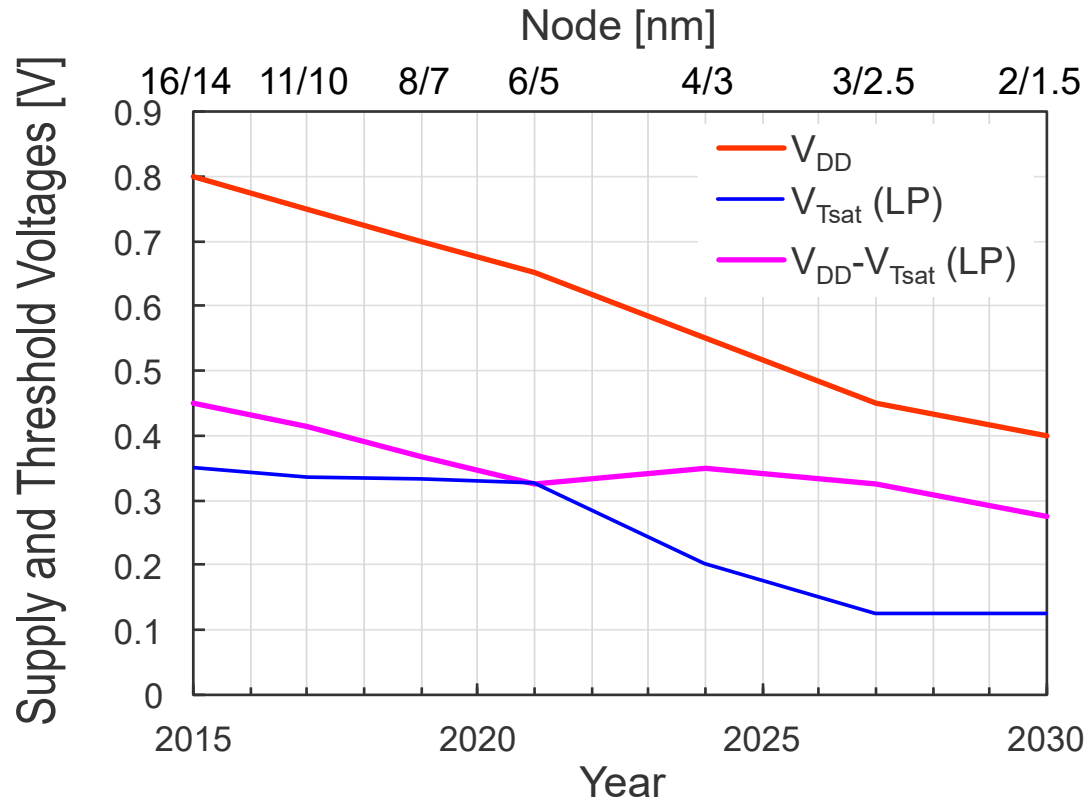
Below 45 nm:
 $0.5V \leq V_{DD} \leq 1V$
 $0.2V \leq V_T \leq 0.5V$

- We clearly have entered the sub-volt era...

The ITRS Roadmap (2011)

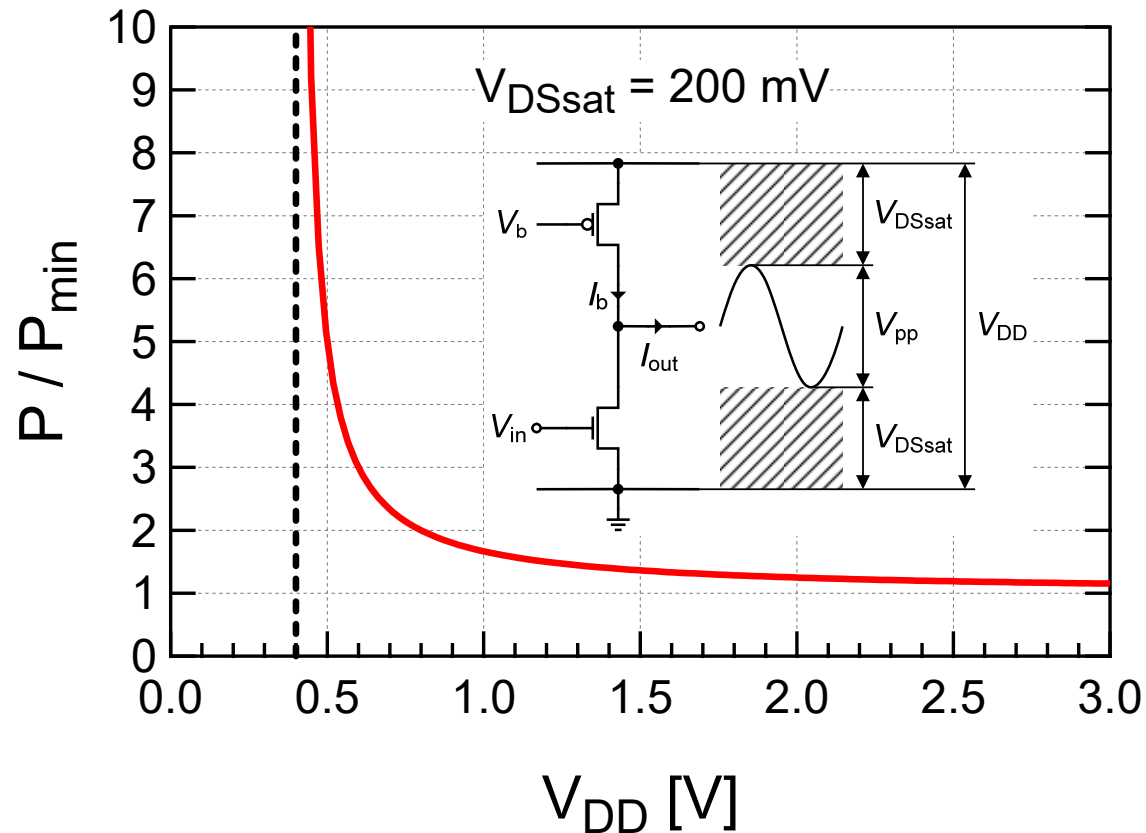


The ITRS Roadmap (2015)



- ...and are heading towards 0.5 V

Voltage Scaling – Against Analog Performance

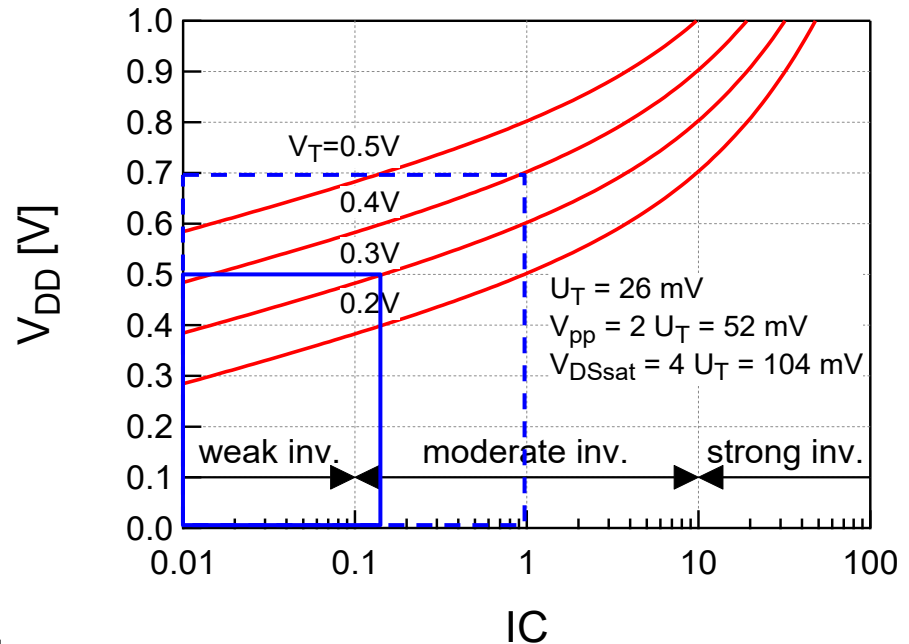
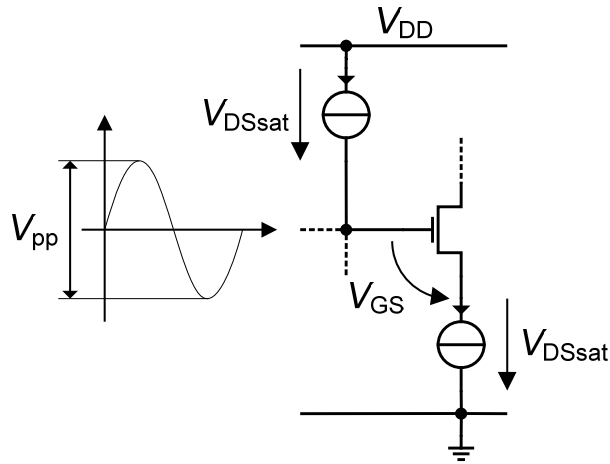


- The power consumption for achieving a **given bandwidth and SNR** is given by

$$\frac{P}{P_{min}} \cong \frac{V_{DD}}{V_{DD} - 2V_{DSsat}}$$

- which becomes very large as V_{DD} gets close to $2V_{DSsat}$

Low-voltage Pushes Bias Points towards Subthreshold



- The supply voltage is set according to

$$V_{DD} = V_{GS} + 2V_{DSsat} + V_{pp}$$

- where V_{GS} depends on the inversion coefficient according to

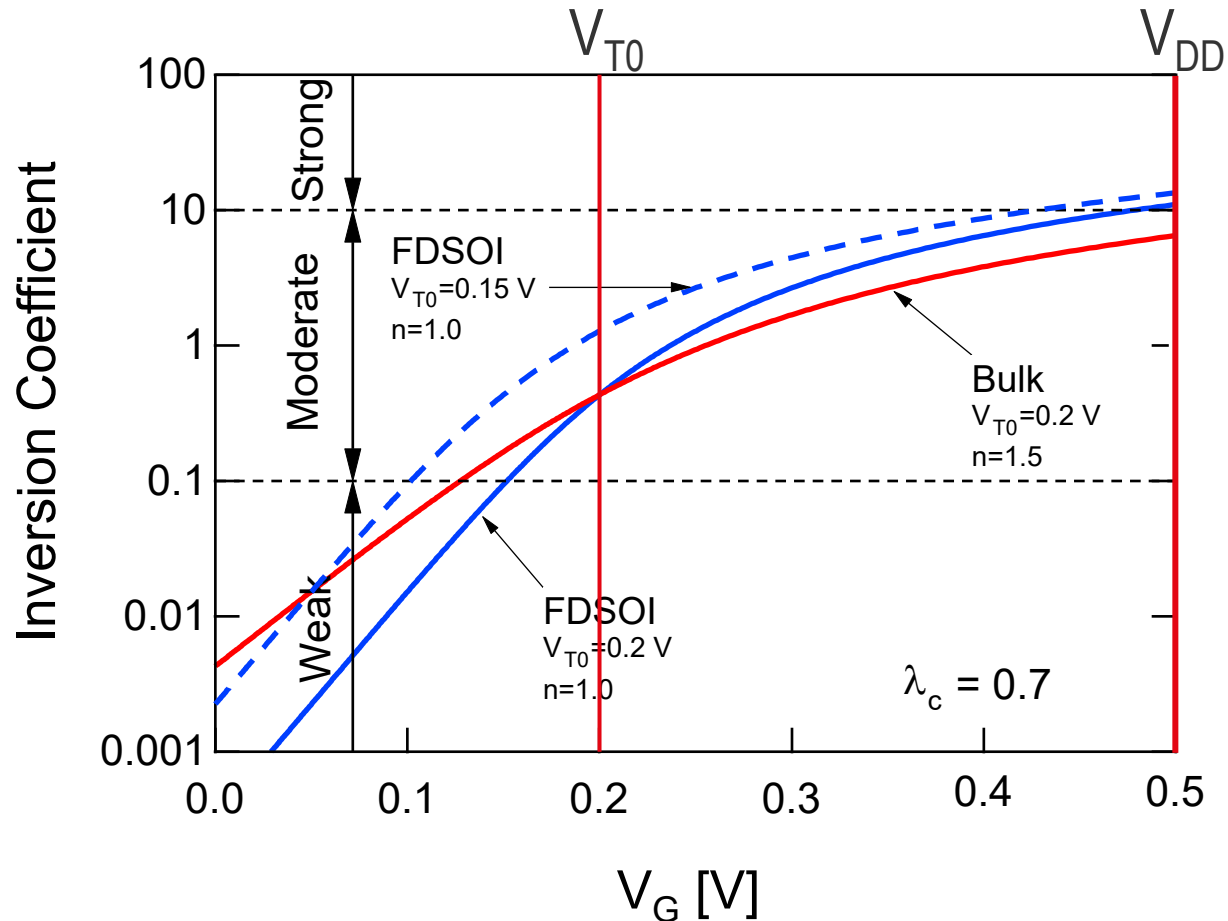
$$V_{GS} = V_T + 2nU_T \ln(e^{\sqrt{IC}} - 1) \text{ with } U_T \triangleq kT/q$$

- Supply voltage below 1V **pushes bias point towards moderate/weak inversion**
- For achieving $V_{DD} = 0.5V$, threshold voltage should be smaller than 0.3 V and bias point has to be in weak inversion ($IC < 0.1$)

Consequences of Supply Voltage Reduction

- Most fundamental
 - ▶ Voltage swing given by $V_{pp} \leq V_{DD} - 2V_{DSSat}$
 - ▶ If $V_{DSSat} = V_{DD}/2$ (or $V_{pp} = 0$) then $K \gg 1$
 - ▶ If voltage is split half between bias and signal: $V_{DSSat} = V_{DD}/4$ (or $V_{pp} = V_{DD}/2$) then increase of K remains acceptable ($K/K_{min} = 2$)
- V_{pp} and V_{DSSat} **must therefore be reduced proportionally with V_{DD}** , consequently, inversion coefficient has to be reduced and **operating point is progressively moving towards moderate and weak inversion**
- Other consequences:
 - ▶ V_{DD} approaching $V_T \rightarrow$ poor switch conductance (eventually conductance gap)
 - ▶ If V_T is lowered \rightarrow open switches start to leak
 - ▶ V_{DD} below $V_{G0} \rightarrow$ requires special band-gap voltage reference circuits
 - ▶ Transistor stacks no more possible \rightarrow requires special LV circuit techniques

Strong Inversion will Disappear at Low-Voltage!



- The above plot clearly illustrates that the strong inversion region is reducing dramatically because of voltage scaling and ultimately is disappearing

Impact of Scaling on Analog/RF Key Parameters

