ELECTRICAL ENGINEERING
UNIVERSITY *of* WASHINGTON

Lecture 14: CMOS Scaling

Based on material prepared by prof. Visvesh S. Sathe

ELECTRICAL & COMPUTER ENGINEERING

Acknowledgements

All class materials (lectures, assignments, etc.) based on material prepared by Prof. Visvesh S. Sathe, and reproduced with his permission

Visvesh S. Sathe
Associate Professor
Georgia Institute of Technology
https://psylab.ece.gatech.edu

UW (2013-2022)
GaTech (2022-present)

Thank you!!

- Teaching this class has been a pleasure ☺

- Feel free to reach out to my personal email: diegopenac94@gmail.com

- Best of luck on your careers! I'm sure you can all become great IC designers. Can't wait to hear of your accomplishments

Some resources

- "The nanosheet transistor is the next (and maybe last) step in Moore's law", Ye et al, IEEE Spectrum, 2019, https://canvas.uw.edu/files/99372084/

- "Nanoscale FinFET Technology for Circuit Designers", by Dr. Alvin Loke - Nov. 2021 (start watching at time 8:16): https://youtu.be/KdBJTqx4Y64?t=496

- Weste & Harris: Section 7.4

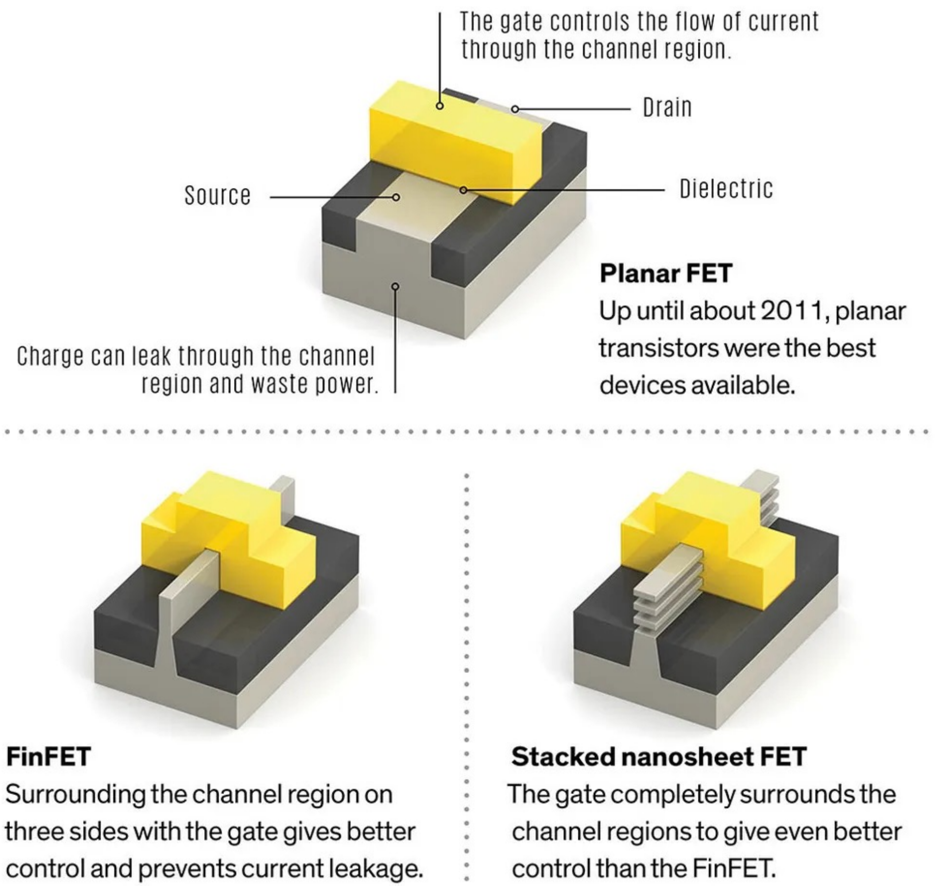
More resources

- Moore’s law:
  - Moore, Gordon E. "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp. 114 ff." IEEE solid-state circuits society newsletter 11.3 (2006): 33-35.

- Dennard’s scaling:
  - R. H. Dennard, F. H. Gaensslen, H. -N. Yu, V. L. Rideout, E. Bassous and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," in IEEE Journal of Solid-State Circuits, vol. 9, no. 5, pp. 256-268, Oct. 1974, doi: 10.1109/JSSC.1974.1050511.

Where is this lecture going?

Evolution of the FET


The gate controls the flow of current through the channel region.
Drain
Source
Dielectric
Charge can leak through the channel region and waste power.

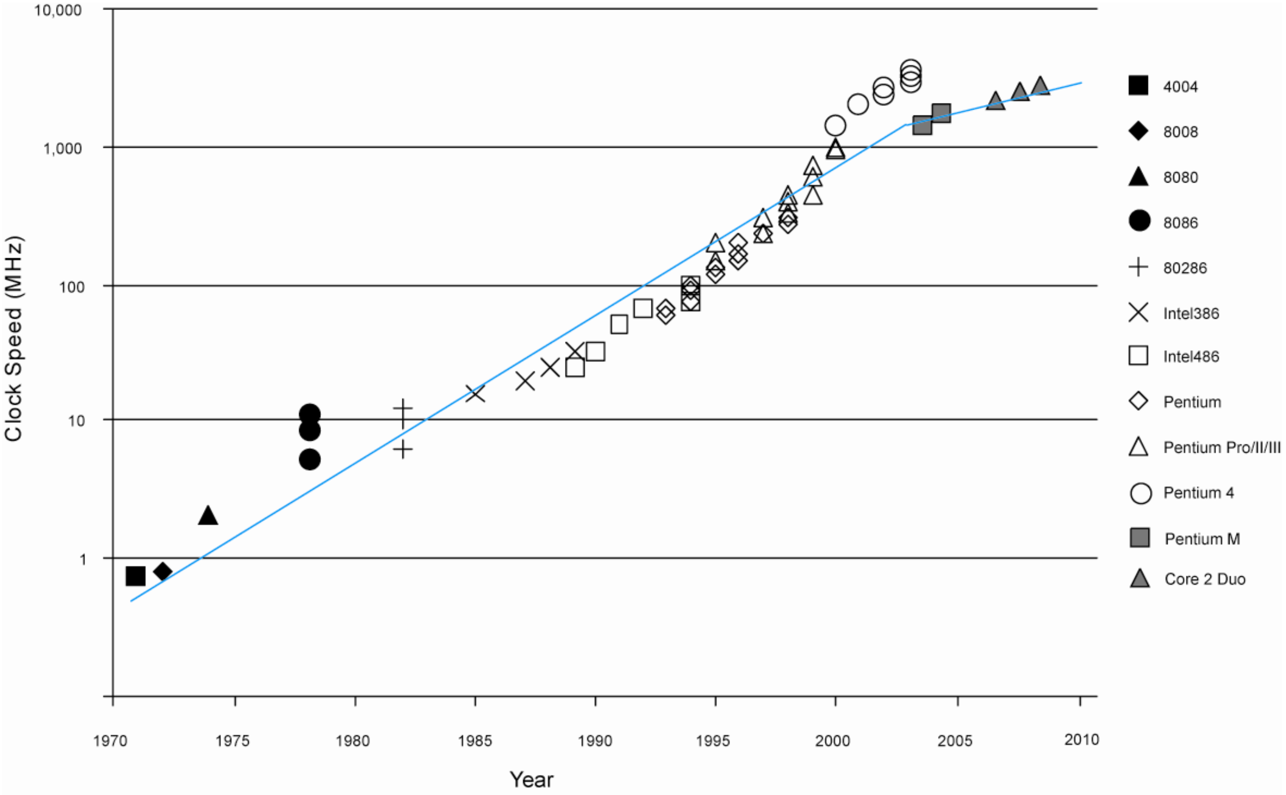
Planar FET
Up until about 2011, planar transistors were the best devices available.

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.

Taken from “The nanosheet transistor is the next (and maybe last) step in Moore’s law”, Ye et al, IEEE Spectrum, 2019

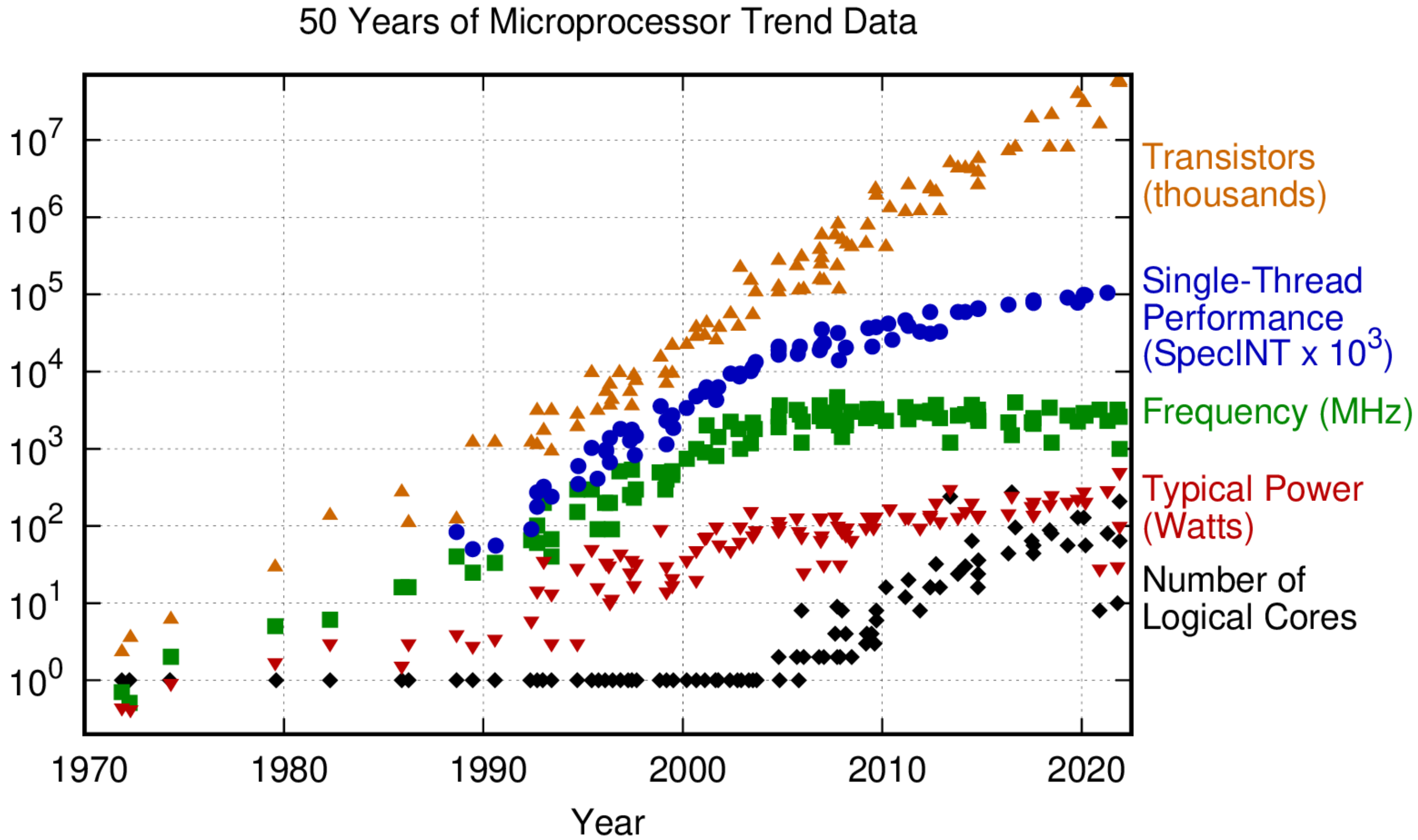
The March of Compute Performance



[Weste,Harris]

- 3 defining trends
  - More “stuff” per mm^2 (Transistors, Wires)
  - Faster transistors → Faster circuits
  - Technology shifts (Much less frequent: Bipolar→NMOS→CMOS →FinFET)

Remember this slide from lecture 1?



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 2010-2021 by K. Rupp

- Taken from https://github.com/karlrupp/microprocessor-trend-data

What happened with voltage scaling?

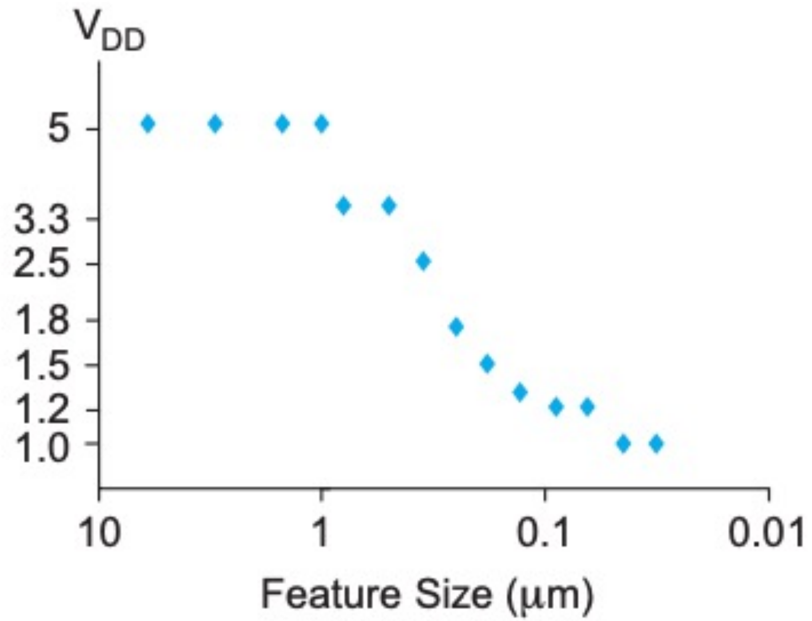


FIGURE 7.14

Voltage scaling with feature size

Taken from W & H

What happened with voltage scaling?

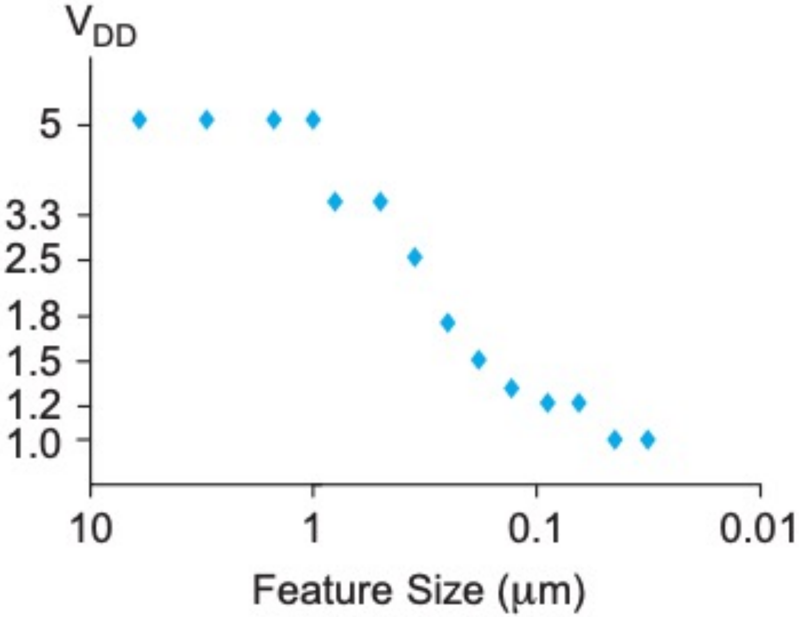


FIGURE 7.14
Voltage scaling with feature size

Taken from W & H

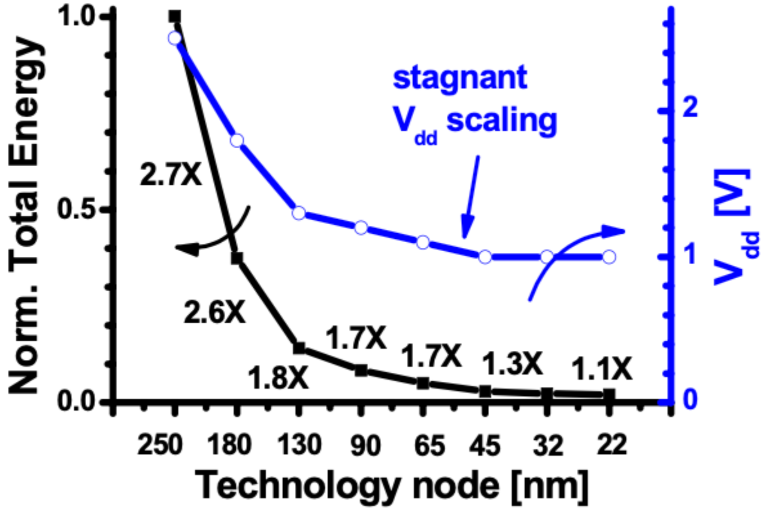


Figure 1: Technology scaling trends of supply voltage and energy.

Taken from: “Near Threshold Computing: Overcoming Performance Degradation from Aggressive Voltage Scaling”, Dreslinski et al, 2009

What happened with voltage scaling?

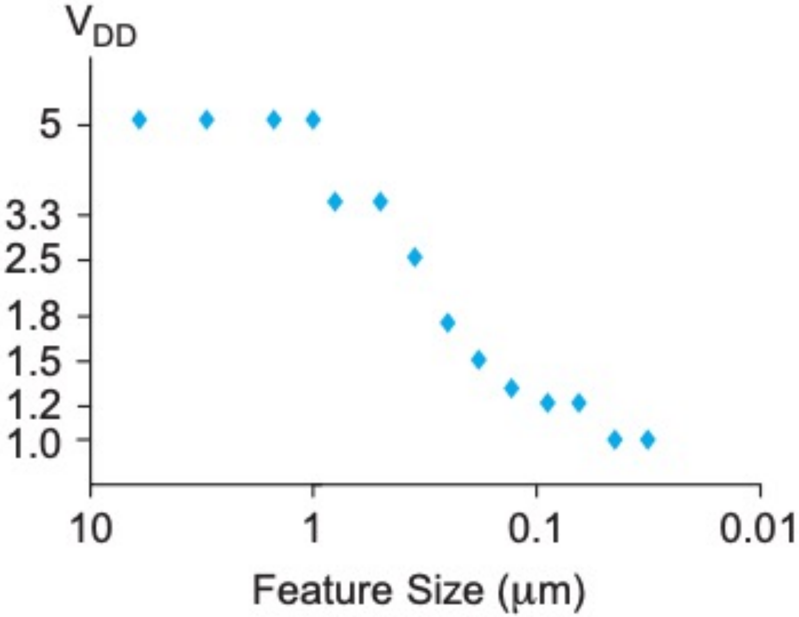


FIGURE 7.14
Voltage scaling with feature size

Taken from W & H

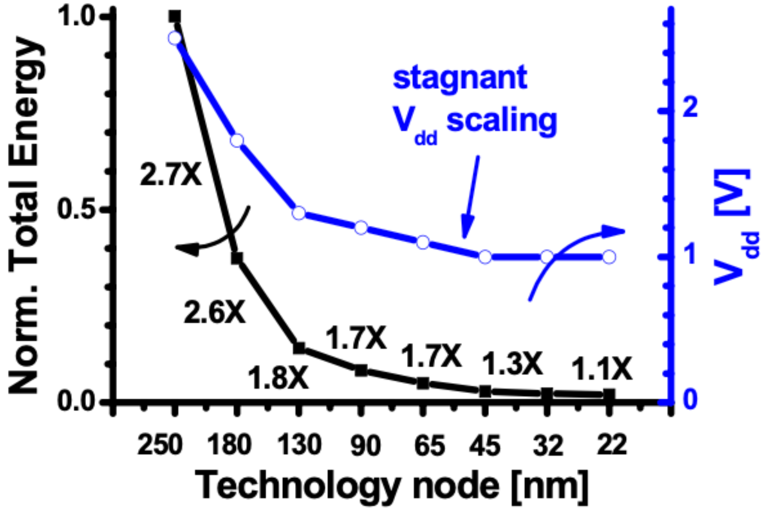


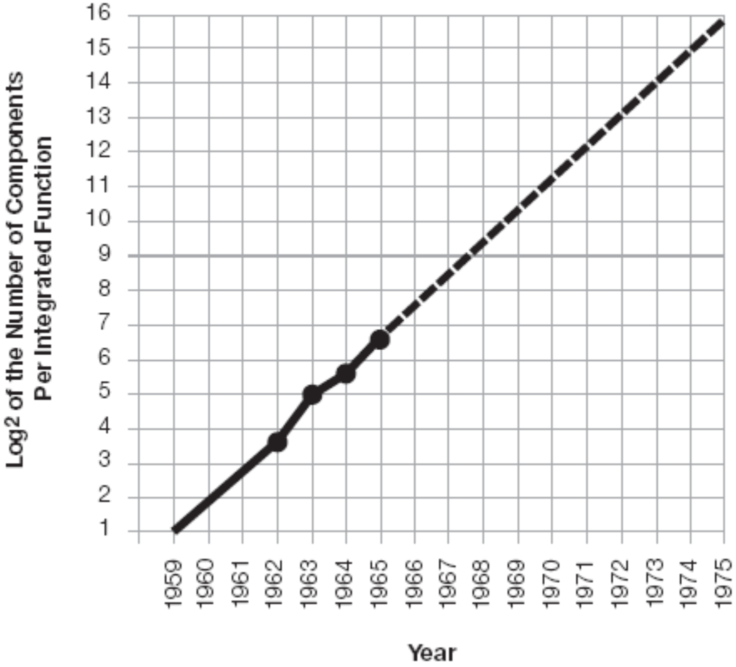
Figure 1: Technology scaling trends of supply voltage and energy.

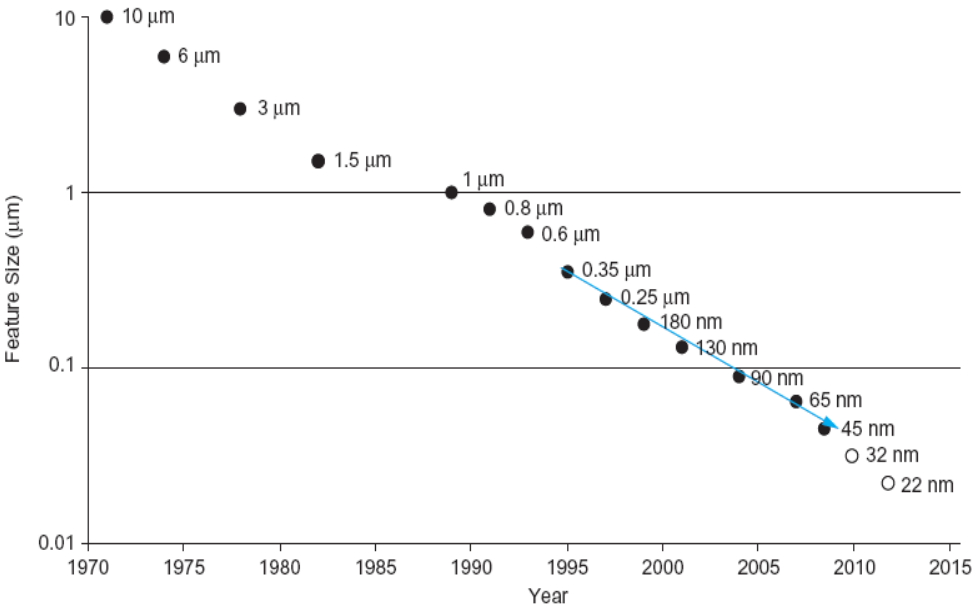
Taken from: “Near Threshold Computing: Overcoming Performance Degradation from Aggressive Voltage Scaling”, Dreslinski et al, 2009

Problem: Vth and leakage current

CMOS Scaling

- Successor to NMOS back in the 1980s : Power Efficiency
- Moore's Law : Driving CMOS scaling over the decades..
  - Transistor densities on a chip 2X every 18** months
  - A statement of fabrication capability, economics
  - More transistors → ↑ functionality, complexity → ↑ compute performance





** Moore himself never claimed 18 months. The original observation was 2X every year and later 2X every 2 years.

CMOS Scaling

- Feature-size shrink by ~30% every 18 months → Smaller transistors
  - Cheaper transistors, more functionality/$
  - Feature size **shrink** factor S=$\sqrt{2}$
  - Each shrink is designated as a new technology “node”

- ↑ transistors/chip does not completely account for ↑ performance
  - Recall that clock frequencies have climbed over the decades
  - Smaller transistors → Faster transistors (Until the last decade)
  - Note : Wires also had to shrink (This will become relevant shortly)

- Dennard’s Law : Smaller transistors = Faster transistors
  - “Constant-field scaling” : Maintain E-field in the transistor across nodes

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	
R: effective resistance	V_{DD}/I_{on}	
C: gate capacitance	WL/t_{ox}	
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	
f: clock frequency	$1/\tau$	
E: switching energy / gate	CV_{DD}^2	
P: switching power / gate	Ef	
A: area per gate	WL	
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	
R: effective resistance	V_{DD}/I_{on}	
C: gate capacitance	WL/t_{ox}	
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	
f: clock frequency	$1/\tau$	
E: switching energy / gate	CV_{DD}^2	
P: switching power / gate	Ef	
A: area per gate	WL	
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	
C: gate capacitance	WL/t_{ox}	
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	
f: clock frequency	$1/\tau$	
E: switching energy / gate	CV_{DD}^2	
P: switching power / gate	Ef	
A: area per gate	WL	
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	
f: clock frequency	$1/\tau$	
E: switching energy / gate	CV_{DD}^2	
P: switching power / gate	Ef	
A: area per gate	WL	
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	1/S
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	
f: clock frequency	$1/\tau$	
E: switching energy / gate	CV_{DD}^2	
P: switching power / gate	Ef	
A: area per gate	WL	
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
    - $1/S^2$ area shrink makes compute cheaper, more powerful
    - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	1/S
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	1/S
f: clock frequency	$1/\tau$	
E: switching energy / gate	CV_{DD}^2	
P: switching power / gate	Ef	
A: area per gate	WL	
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	1/S
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	1/S
f: clock frequency	$1/\tau$	**S**
E: switching energy / gate	CV_{DD}^2	
P: switching power / gate	Ef	
A: area per gate	WL	
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	1/S
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	1/S
f: clock frequency	$1/\tau$	**S**
E: switching energy / gate	CV_{DD}^2	**$1/S^3$**
P: switching power / gate	Ef	
A: area per gate	WL	
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	1/S
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	1/S
f: clock frequency	$1/\tau$	**S**
E: switching energy / gate	CV_{DD}^2	**$1/S^3$**
P: switching power / gate	Ef	**$1/S^2$**
A: area per gate	WL	
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	1/S
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	1/S
f: clock frequency	$1/\tau$	**S**
E: switching energy / gate	CV_{DD}^2	**$1/S^3$**
P: switching power / gate	Ef	**$1/S^2$**
A: area per gate	WL	**$1/S^2$**
Switching power density	P/A	
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	1/S
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	1/S
f: clock frequency	$1/\tau$	**S**
E: switching energy / gate	CV_{DD}^2	**$1/S^3$**
P: switching power / gate	Ef	**$1/S^2$**
A: area per gate	WL	**$1/S^2$**
Switching power density	P/A	**1**
Switching current density	I_{on}/A	

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
  - $1/S^2$ area shrink makes compute cheaper, more powerful
  - BUT must be accompanied with $1/S^2$ power savings!!!

Constant Field Scaling : Impact on Transistors

Parameter	Note	Dennard Scaling
L: Length	Enabled by Fab.	1/S
W: Width	Enabled by Fab.	1/S
t_{ox}: gate oxide thickness	Enabled by Fab.	1/S
V_{DD}: supply voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	1/S
V_t: threshold voltage	$E_{ox}=(V_{DD}-V_{th})/C_{ox}$	**1/S**
NA: substrate doping	**Optional**	S
β	$W\varepsilon/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	1/S
τ: gate delay	$CV_{DD}/\beta(V_{DD}-V_t)^2$	1/S
f: clock frequency	$1/\tau$	**S**
E: switching energy / gate	CV_{DD}^2	**$1/S^3$**
P: switching power / gate	Ef	**$1/S^2$**
A: area per gate	WL	**$1/S^2$**
Switching power density	P/A	**1**
Switching current density	I_{on}/A	S

[Adapted from Weste,Harris]

- $S = \sqrt{2}$
- Device current
- But so does V_{dd}, C^*_{load}
- Overall delay
- Power
- Power Density:
    - $1/S^2$ area shrink makes compute cheaper, more powerful
    - BUT must be accompanied with $1/S^2$ power savings!!!

Think about this at home

- Ring oscillator (Assume load presented only by devices):
  - Case1: Only Width of devices scales by 1/S
  - Case2: All tech. parameters scale (W, L, V, t_{ox} , V_{th}) all scale by a factor of 1/S .
  - Case3: Width and Length of devices scales by 1/S
  - What is change in delay, power, energy-per-cycle?
- FETs are all held at 1 micron after scaling to the new technology generation….what is the change in delay, power, energy..

Interconnect Scaling (think about this at home)

Parameter	Sensitivity	Scale Factor
w: width		**1/S**
s: spacing		**1/S**
t: thickness*		1/S*
h: height		1/S
D_c: die size		D_c
R_w: wire resistance/unit length	1/wt	S^2
C_{wf}: fringing capacitance / unit length	t/s	1
C_{wp}: parallel plate capacitance / unit length	w/h	1
C_w: total wire capacitance / unit length	$C_{wf} + C_{wp}$	1
t_{wu}: unrepeated RC delay / unit length	R_wC_w	S^2
t_{wr}: repeated RC delay / unit length	$\text{sqrt}(RCR_wC_w)^*$	sqrt(S)
Crosstalk noise	w/h	1
E_w: energy per bit / unit length	$C_wV_{DD}^2$	$1/S^2$

[Weste,Harris]

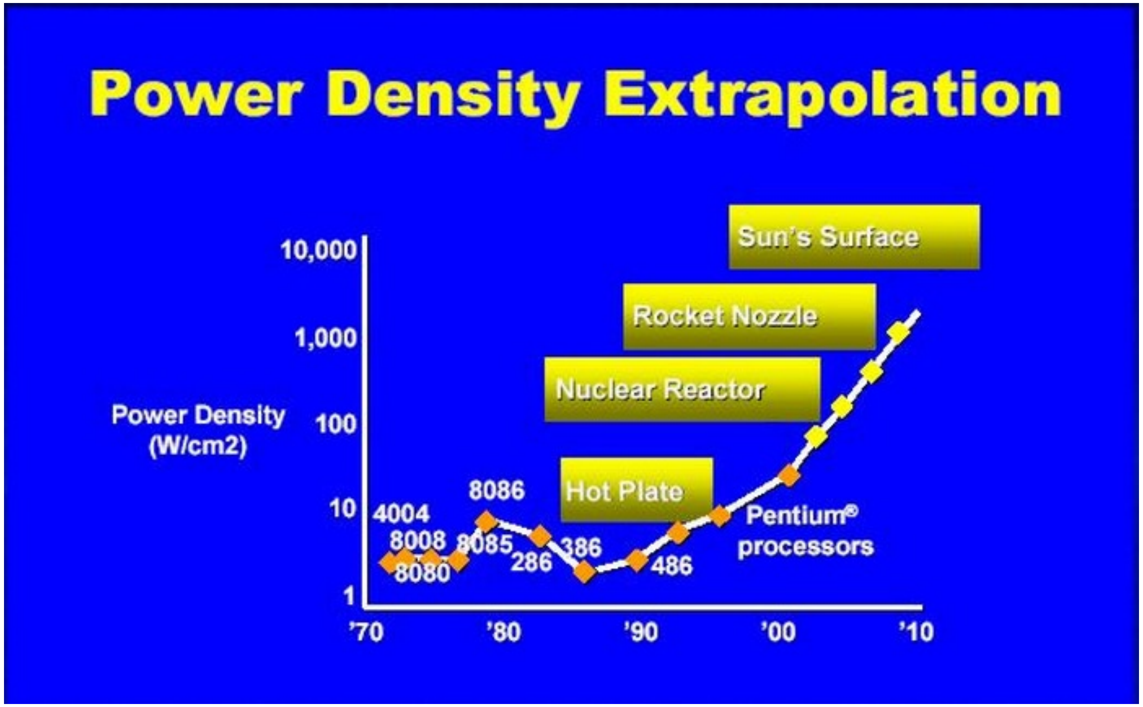
- Interconnect MUST scale along with transistors (Why?)
- *Thickness: To scale or not to scale…
  - Scale it, and R↑ as S^2, RC as S^2.
  - Leave it unscaled, R↑ S, but RC still as S^{2**}. Fabrication challenge

Moore and Dennard Laws Part Ways…

Parameter	Dennard Scaling	Reality
L: Length	1/S	1/S
W: Width	1/S	1/S
t_{ox}: gate oxide thickness	1/S	**~1/S**
V_{DD}: supply voltage	**1/S**	**1**
V_t: threshold voltage	**1/S**	**1**
β ($\mu C_{ox}W/L$)	**S**	**1**
I_{on}: ON current	1/S	1/S
R: effective resistance	1	1
Gate + Wire Capacitance	1/S	1< x <1/S
τ: gate delay	1/S	1< x <1/S
f: clock frequency	**S**	**<S**
E: switching energy / gate	**$1/S^3$**	**>1/S**
P: switching power / gate	**$1/S^2$**	**>1/S**
A: area per gate	$1/S^2$	$1/S^2$
Switching power density	**1**	**>S**
Switching current density	S	S

- Velocity saturation
- V_{th} scaling ended
- V_{dd} scaling ended
- ↑ Wire parasitics
- t_{ox} scaling limited by gate leakage

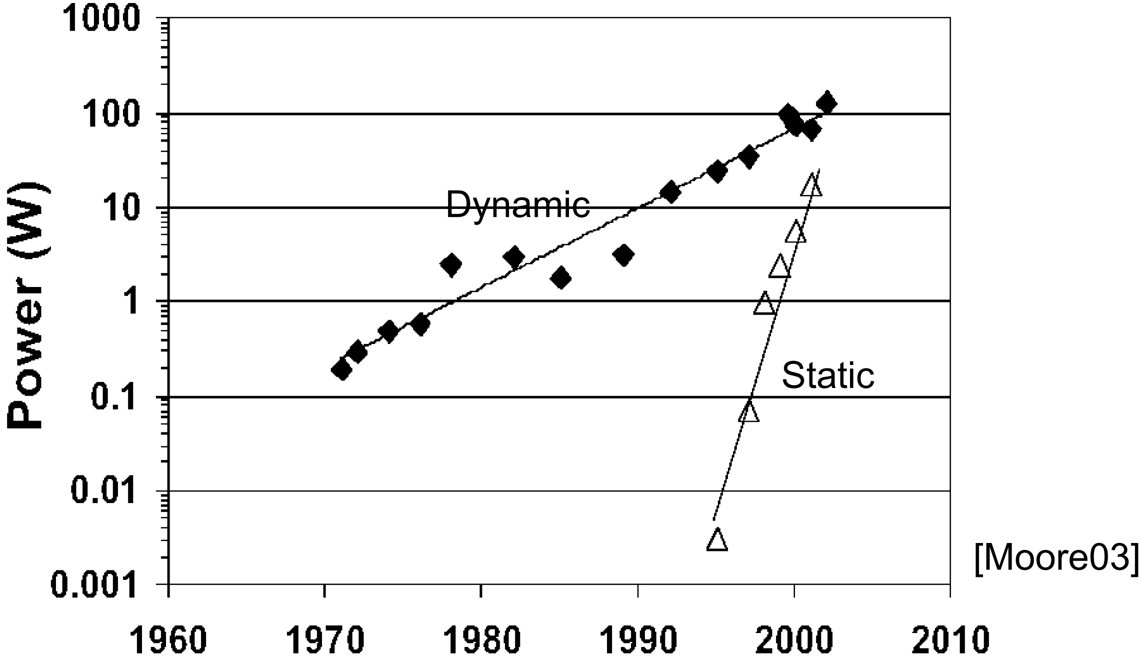
Dynamic Power



[Intel]

- Patrick Gelsinger (now CEO of Intel) @ISSCC 2001
  - Current trend ➔ Power density comparable to the Sun’s surface by 2010
- That did not happen of course (Frequency flattened, Power Mgmt.)
- Energy efficient design emerged as a key limiter in early 2000s

V_{th} Scaling



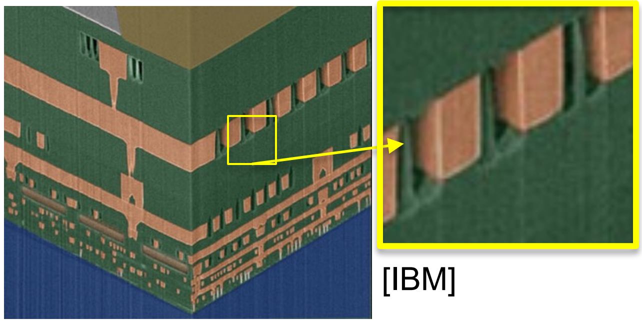
Static Power: Subthreshold (V_{th}) + Gate (t_{ox}) leakage

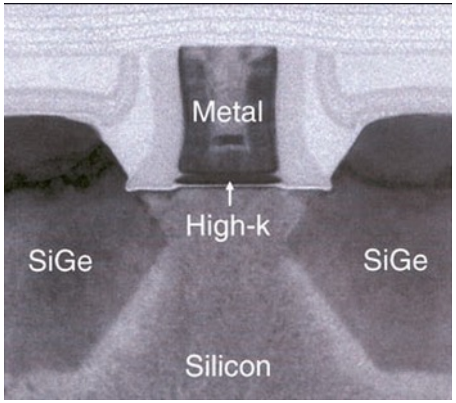
- V_{th} scaling worked well for the longest time
  - $P_{leak} = ke^{-Vth}$ but $P_{leak} << P_{dynamic}$ it did not matter….till recently
- Gate leakage (e^{-tox}) also began to ↑. Prevented t_{ox} reduction
- Inability to ↓ V_{th} crucial for improved device performance
- Contributed to the end of V_{dd} scaling ➔ the ensuing efficiency crisis

(Relatively) Recent Developments

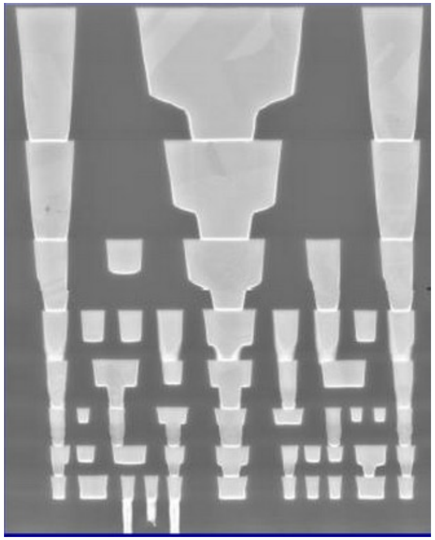
- Dennard's law essentially died ~ early 2000s
- Process technology kept Moore's law going
- Back End Technology advances
  - Copper Interconnect
  - Low-K Dielectrics (Lower wire capacitance)
  - Tapered Back-End stack (Global Interconnect)

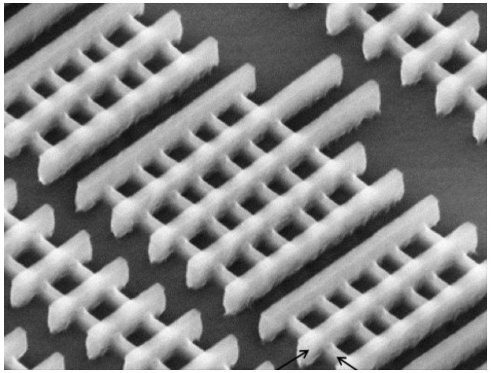
- Front-End technology advances
  - Strained Silicon (Enhance μ)
    - Dual Stress Liners
    - eSiGe
  - High-K Dielectrics
  - Metal gate
  - Tri-gate MOSFETS


[IBM]



[Intel]

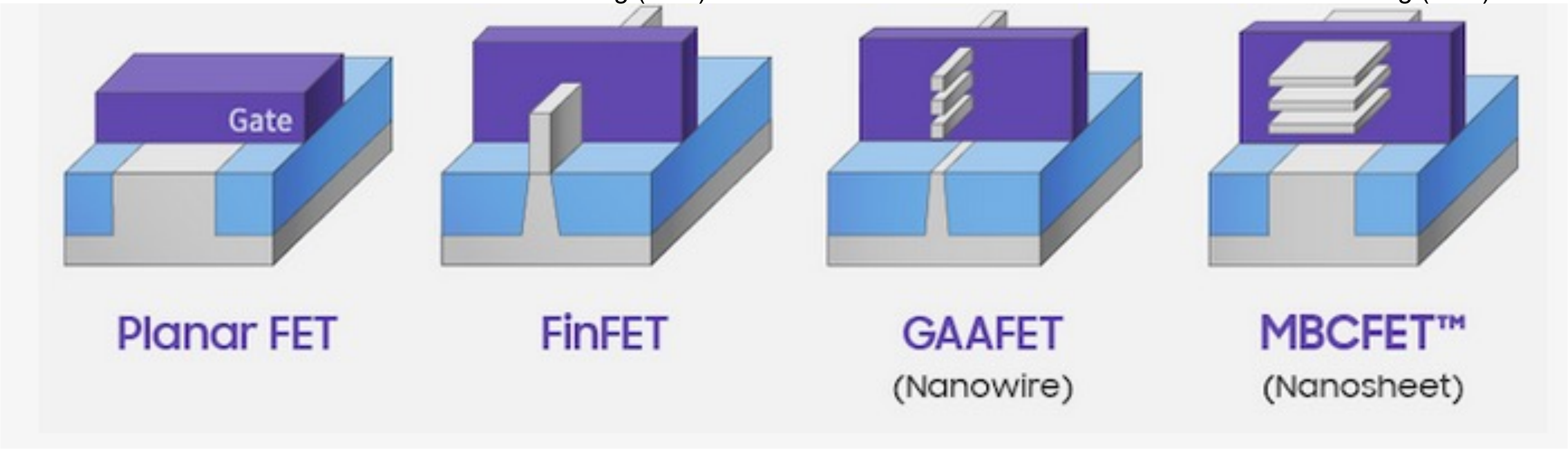

[Intel]


[Intel]

Device Architecture Evolution…

Samsung (5nm)

Samsung (3nm)

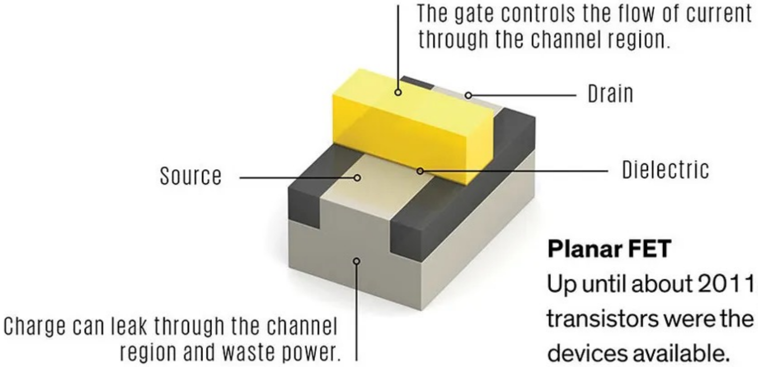


Source: Anandtech

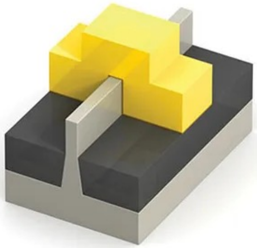
- Continued push:
  - scaled geometries
  - improved short-channel performance (How well does my gate control the FET)

Device Architecture Evolution… (continued)

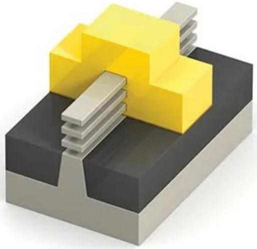
Evolution of the FET


The gate controls the flow of current through the channel region.
Drain
Source
Dielectric
Charge can leak through the channel region and waste power.

Planar FET
Up until about 2011, planar transistors were the best devices available.

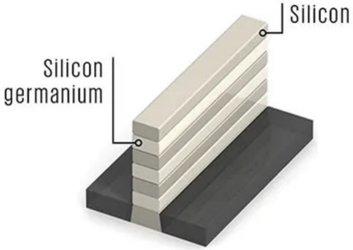


FinFET
Surrounding the channel region on three sides with the gate gives better control and prevents current leakage.

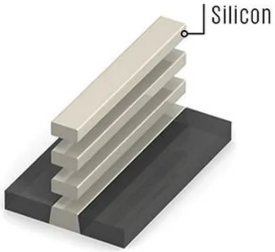


Stacked nanosheet FET
The gate completely surrounds the channel regions to give even better control than the FinFET.

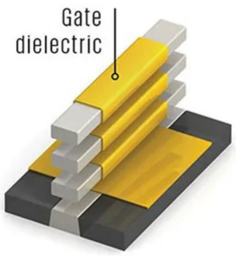
How to make nanosheets?



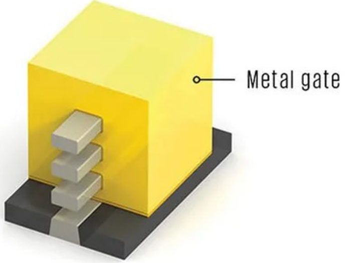
A superlattice of silicon and silicon germanium are grown atop the silicon substrate.



A chemical that etches away silicon germanium reveals the silicon channel regions.



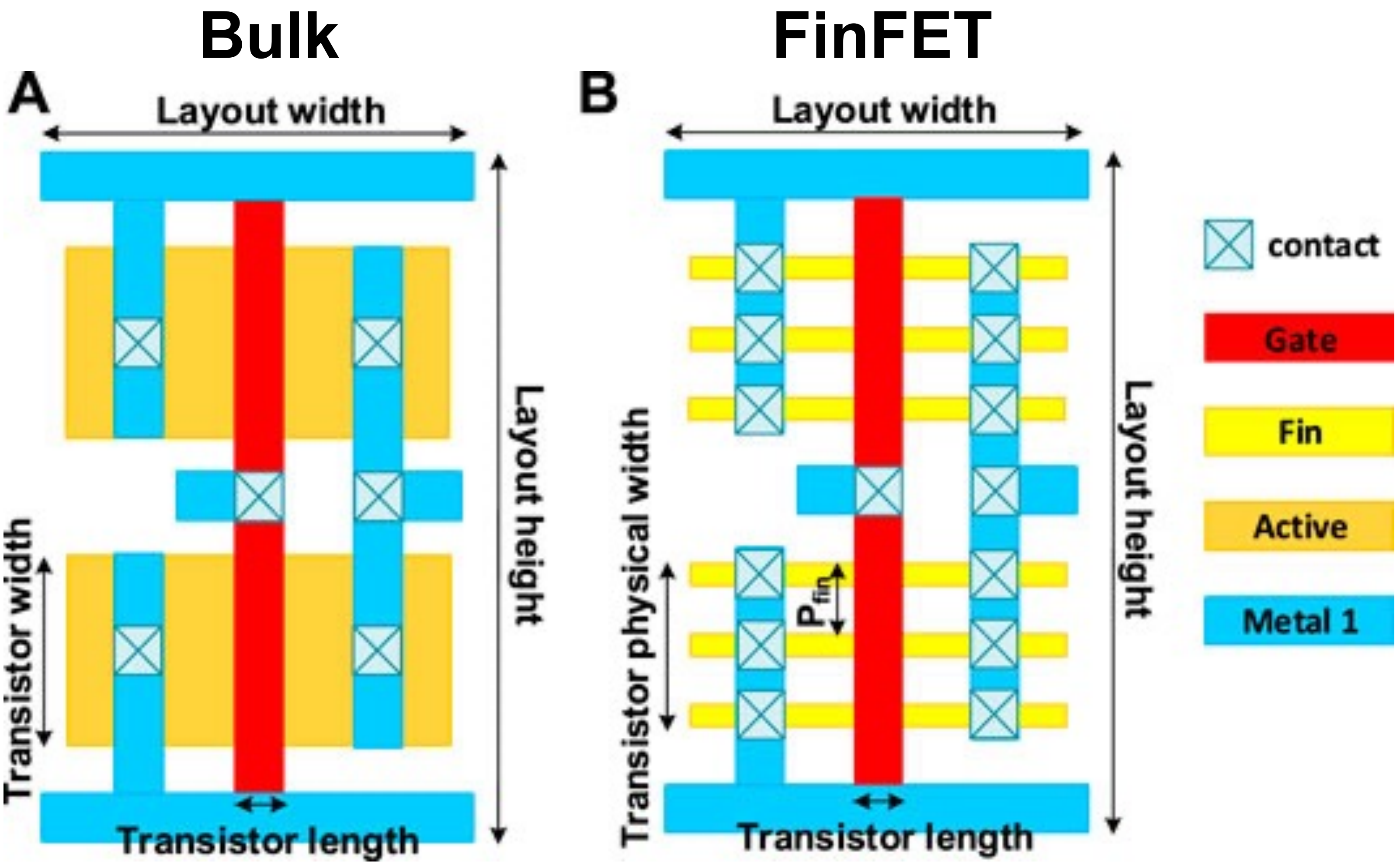
Atomic layer deposition builds a thin layer of dielectric on the silicon channels, including on the underside.



Atomic layer deposition builds the metal gate so that it completely surrounds the channel regions.

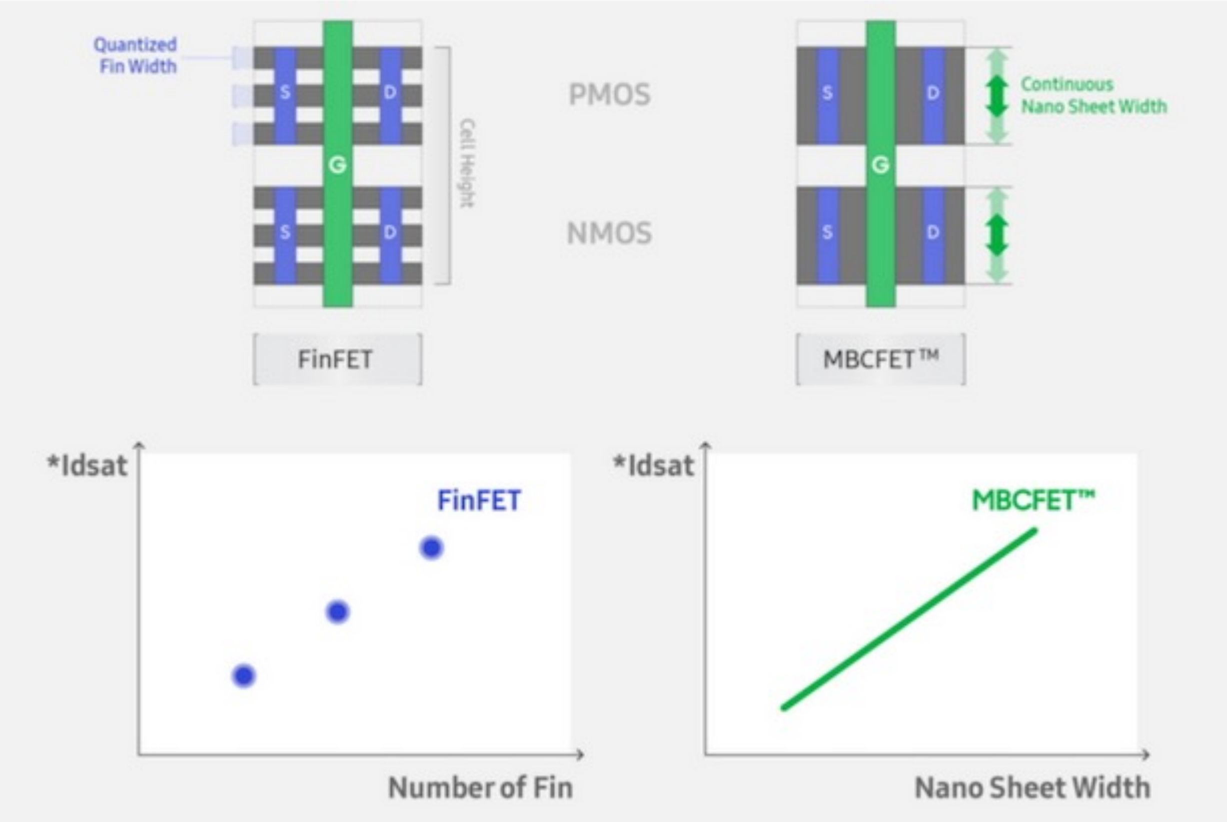
Taken from “The nanosheet transistor is the next (and maybe last) step in Moore’s law”, Ye et al, IEEE Spectrum, 2019

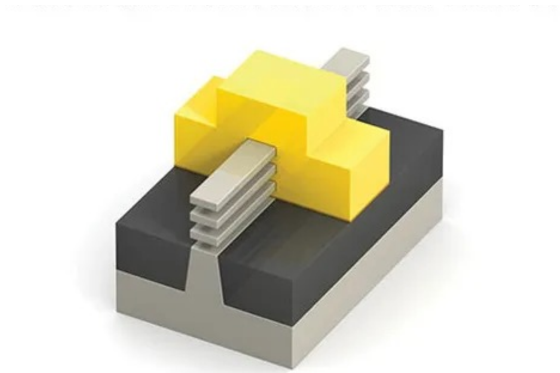
An inverter using FinFETs



Taken from Lu, Anni, et al. "NeuroSim simulator for compute-in-memory hardware accelerator: validation and benchmark." Frontiers in Artificial Intelligence (2021): 70.

Nanosheet-based FET benefits





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

- Improved short-channel properties (gate all around)
- More gate width flexibility (more sheet widths possible)