CMOS Scaling

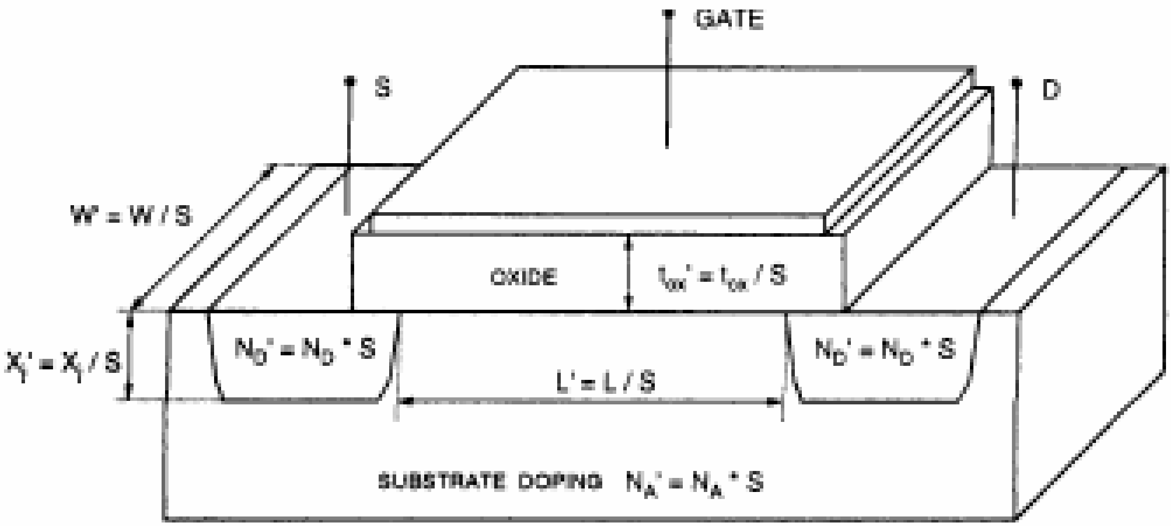
Two motivations to scale down

Faster transistors, both digital and analog

To pack more functionality per area.

Lower the cost!

Full Scaling (Constant-Field Scaling) (which makes (some) physical sense)



Scale all dimensions and voltages by the same factor, so the field does not change, and the device physics will be about the same. Makes sense at the first glance.

Table 3.2 Full scaling of MOSFET dimensions, potentials, and doping densities

Quantity	Before scaling	After scaling
Channel length	L	$L' = L/S$ $(S > 1)$
Channel width	W	$W' = W/S$
Gate oxide thickness	t_{ox}	$t'_{ox} = t_{ox}/S$
Junction depth	x_j	$x'_j = x_j/S$
Power supply voltage	V_{DD}	$V'_{DD} = V_{DD}/S$
Threshold voltage	V_{T0}	$V'_{T0} = V_{T0}/S$
Doping densities	N_A	$N'_A = S \cdot N_A$
	N_D	$N'_D = S \cdot N_D$

Will talk about these later

Constant filed ➔ constant current density (A/cm)

$W' = W/S$ ⇨ $I_D' = I_D/S$ (S > 1)

$$P = I_D \cdot V_{DS}$$

(holds for every instant)

$$P' = I_D' \cdot V_{DS}' = \frac{1}{S^2} \cdot I_D \cdot V_{DS} = \frac{P}{S^2}$$

Device resistance $R = V/I$

Load capacitance $C = C_{ox}WL$. $C' = C/S$ Why? And why not so true?

Switch delay time $\tau \sim RC$. $\tau' = \tau/S$

$f_T \propto (V_G - V_{th})/L^2$ or $f_T \propto v_{sat}/L$ How's f_T doing?

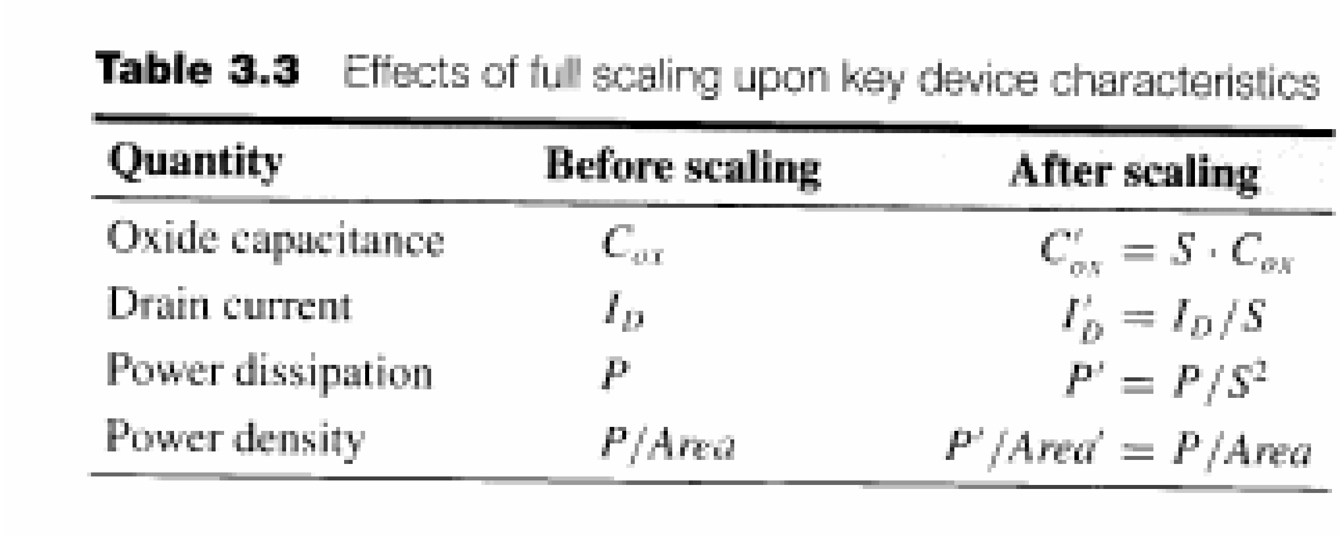
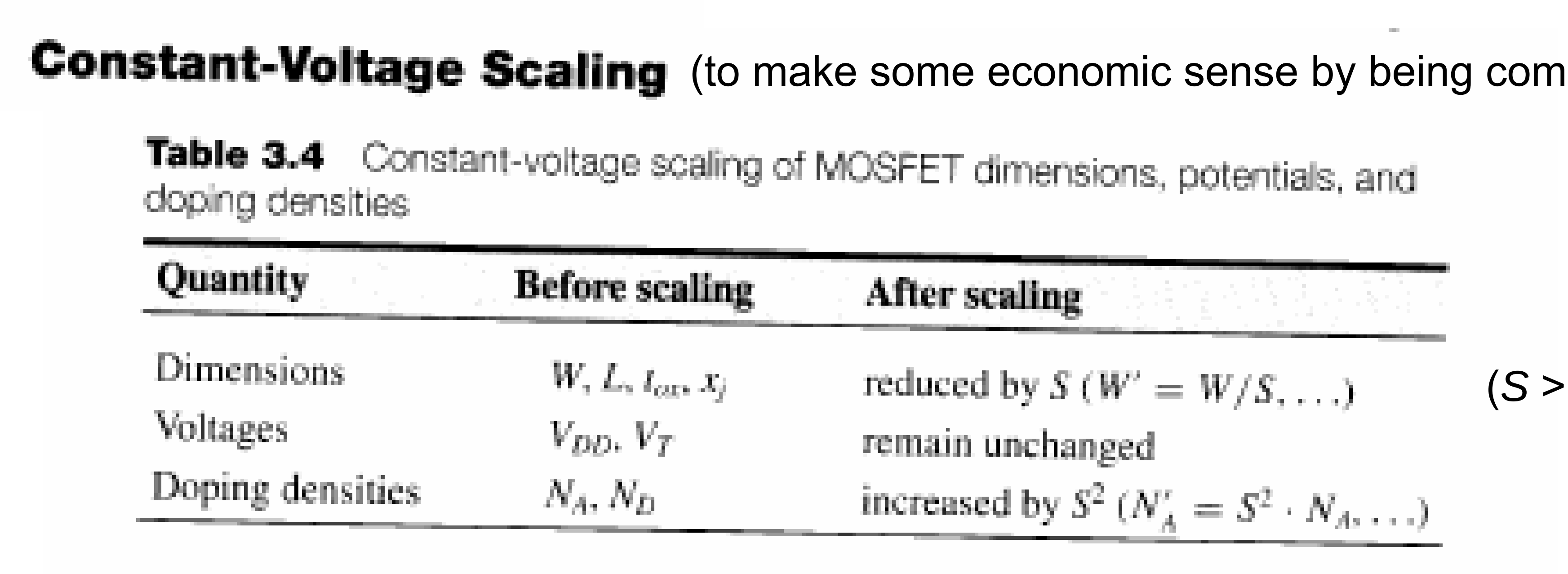


Table 3.3 Effects of full scaling upon key device characteristics

Quantity	Before scaling	After scaling
Oxide capacitance	C_{ox}	$C'_{ox} = S \cdot C_{ox}$
Drain current	I_D	$I'_D = I_D/S$
Power dissipation	P	$P' = P/S^2$
Power density	$P/Area$	$P'/Area' = P/Area$



Constant-Voltage Scaling (to make some economic sense by being compatible)

Table 3.4 Constant-voltage scaling of MOSFET dimensions, potentials, and doping densities

Quantity	Before scaling	After scaling
Dimensions	W, L, t_{ox}, x_j	reduced by S ($W' = W/S, \ldots$)
Voltages	V_{DD}, V_T	remain unchanged
Doping densities	N_A, N_D	increased by S^2 ($N_A' = S^2 \cdot N_A, \ldots$)

($S > 1$)

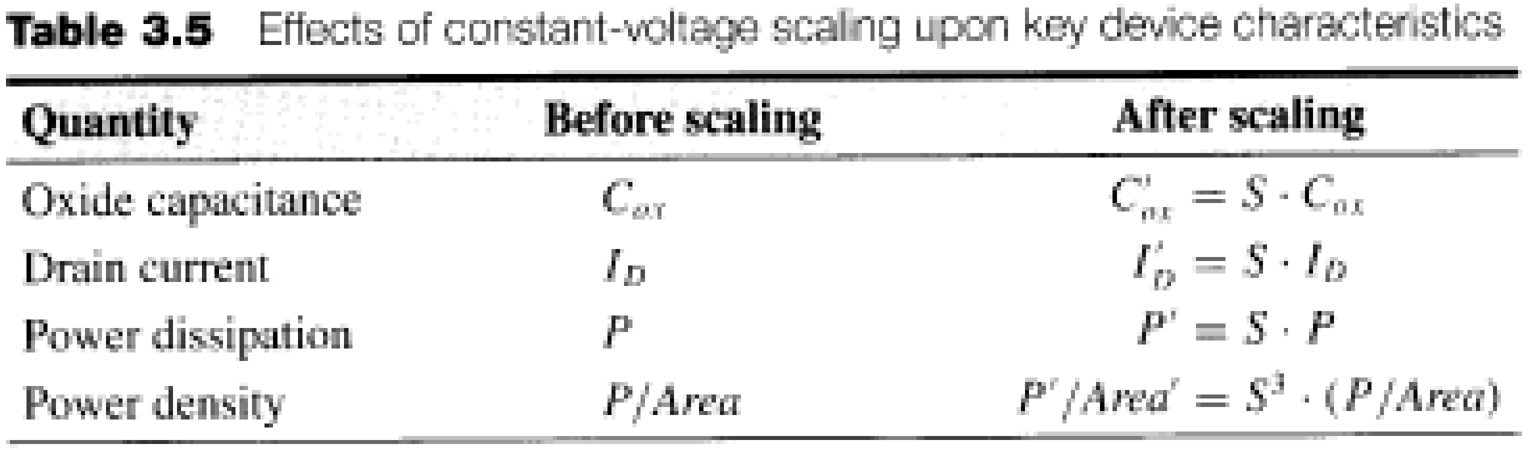


Table 3.5 Effects of constant-voltage scaling upon key device characteristics

Quantity	Before scaling	After scaling
Oxide capacitance	C_{ox}	$C_{ox}' = S \cdot C_{ox}$
Drain current	I_D	$I_D' = S \cdot I_D$
Power dissipation	P	$P' = S \cdot P$
Power density	$P/Area$	$P'/Area' = S^3 \cdot (P/Area)$

Switching delay time $1/S^2$

Moore’s Law

- Gordon E. Moore (born 1929), a co-founder of Intel

- Moore's Law is the **empirical observation** made in 1965 that the number of transistors on an integrated circuit **for minimum component cost** doubles every 24 months (sometimes quoted as 18 months).

- Moore's law is not about just the density of transistors that can be achieved, but about the density of transistors at which the cost per transistor is the lowest.

Why Small (if they are already so small that we can’t make them better)?

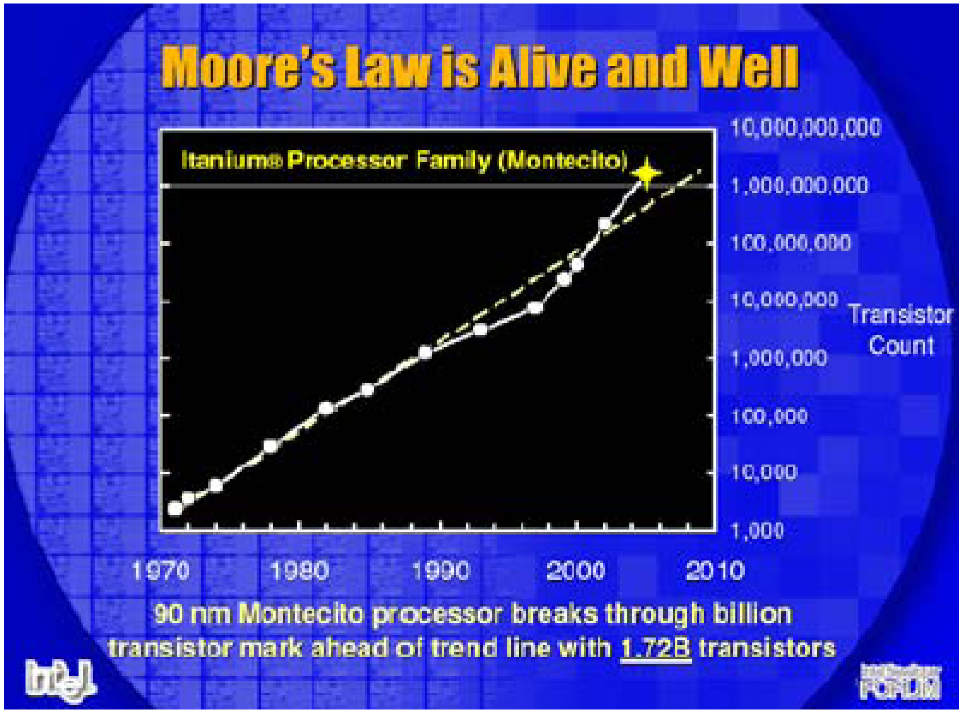
- To pack more functionality into each unit area
  - With feature size shrinking and wafer size growing, more and better products each wafer – better economics

Transistor radio!

1965: $1/transistor

2006: 1 “microdolar”/transistor

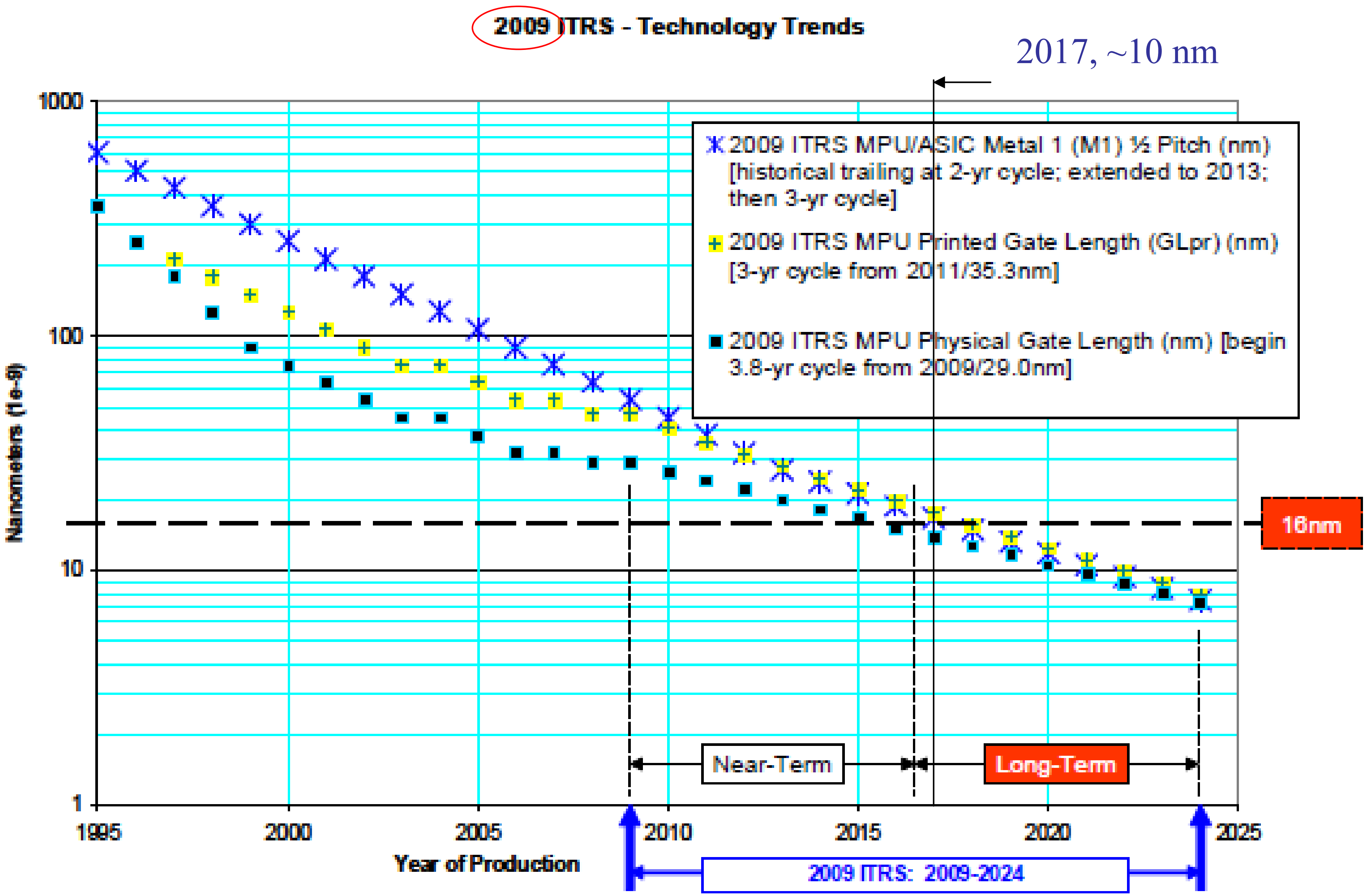
Moore’s Law is Alive and Well



Moore’s Law is Alive and Well

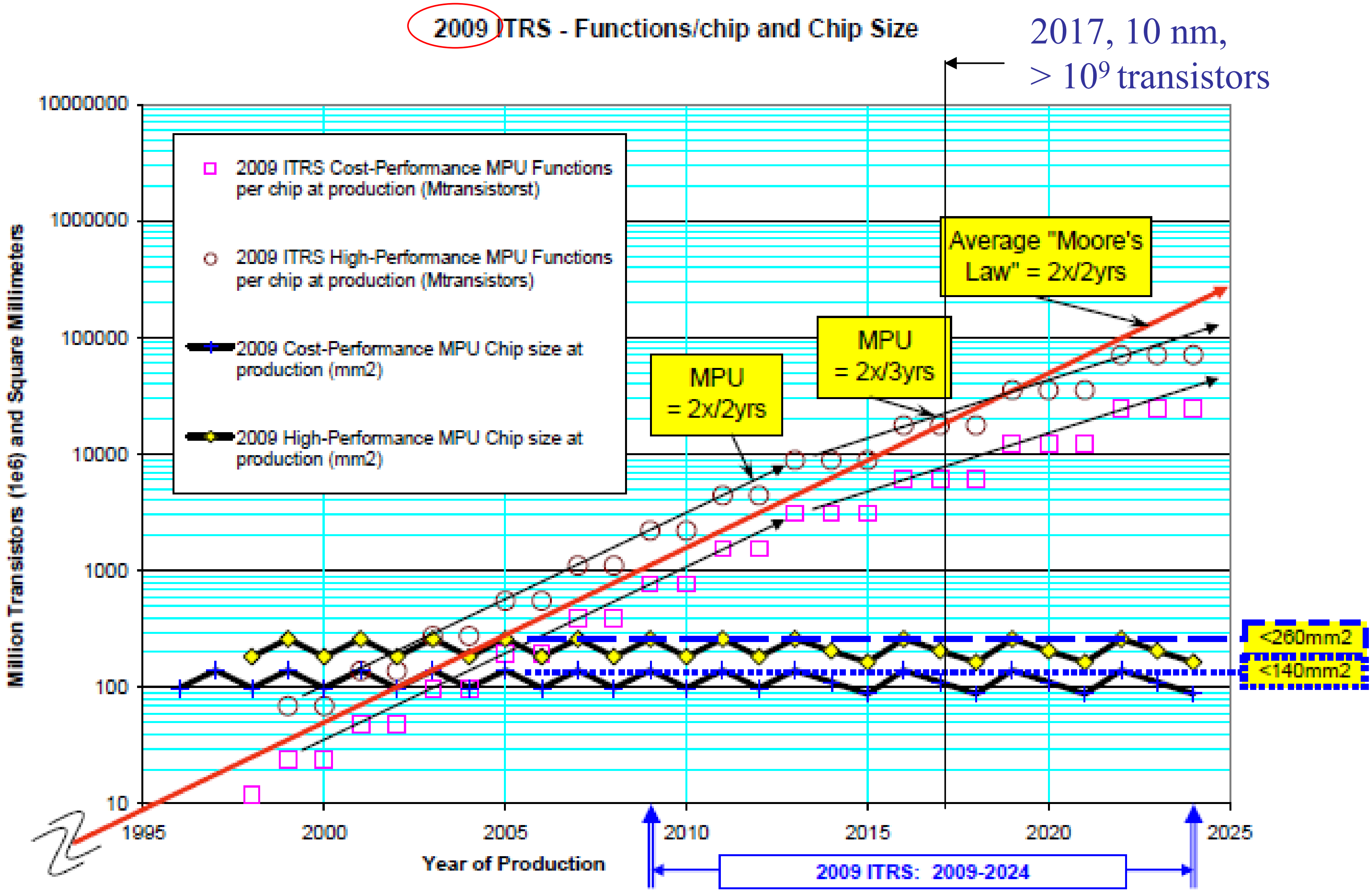
90 nm Montecito processor breaks through billion transistor mark ahead of trend line with 1.72B transistors

Source: M. Bohr, Intel Development Forum, September 2004.


2009 ITRS - Technology Trends
1000
100
10
1
Nanometers (1e-9)
1995
2000
2005
2010
2015
2020
2025
Year of Production
2009 ITRS MPU/ASIC Metal 1 (M1) ½ Pitch (nm) [historical trailing at 2-yr cycle; extended to 2013; then 3-yr cycle]
2009 ITRS MPU Printed Gate Length (GLpr) (nm) [3-yr cycle from 2011/35.3nm]
2009 ITRS MPU Physical Gate Length (nm) [begin 3.8-yr cycle from 2009/29.0nm]
16nm
Near-Term
Long-Term
2009 ITRS: 2009-2024

2017, ~10 nm

With 10 nm feature size, we are near the end of the road...



Can Moore’s Law Go On Forever?

- People have had doubts for many years

- But so far so good (?)

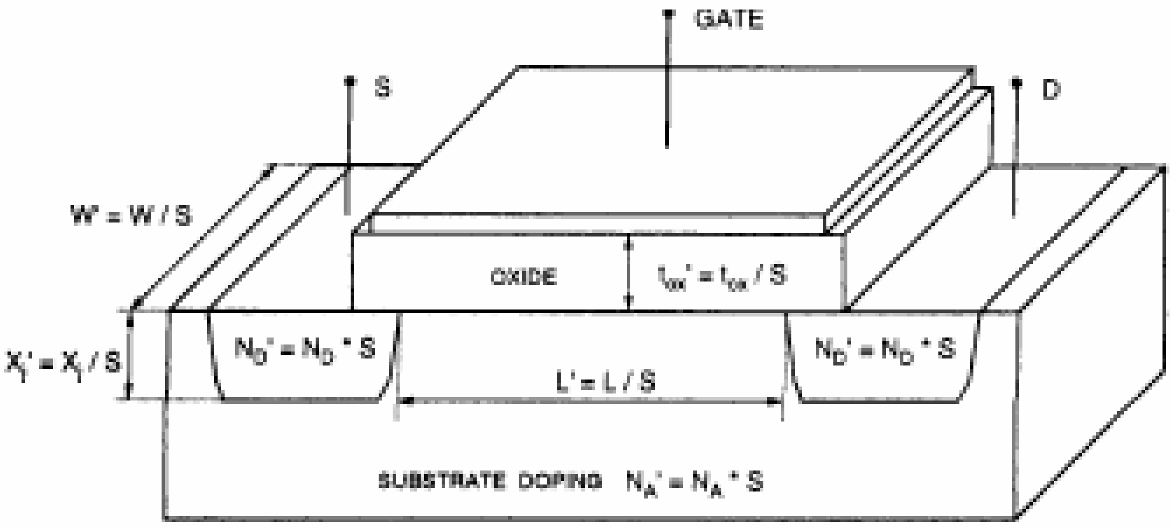
- In the past, the doubts were more about processing limits than physics limits

- Processing limits have been overcome
  And probably will continue to be overcome

- Are we hitting the physics limits?

Before we look at the physics limits and alternative materials/devices, let’s look at issues with CMOS scaling and new CMOS structures that currently let CMOS get by.

Full Scaling (Constant-Field Scaling) (which makes (some) physical sense)



Scale all dimensions and voltages by the same factor, so the field does not change, and the device physics will be about the same. Makes sense at the first glance.

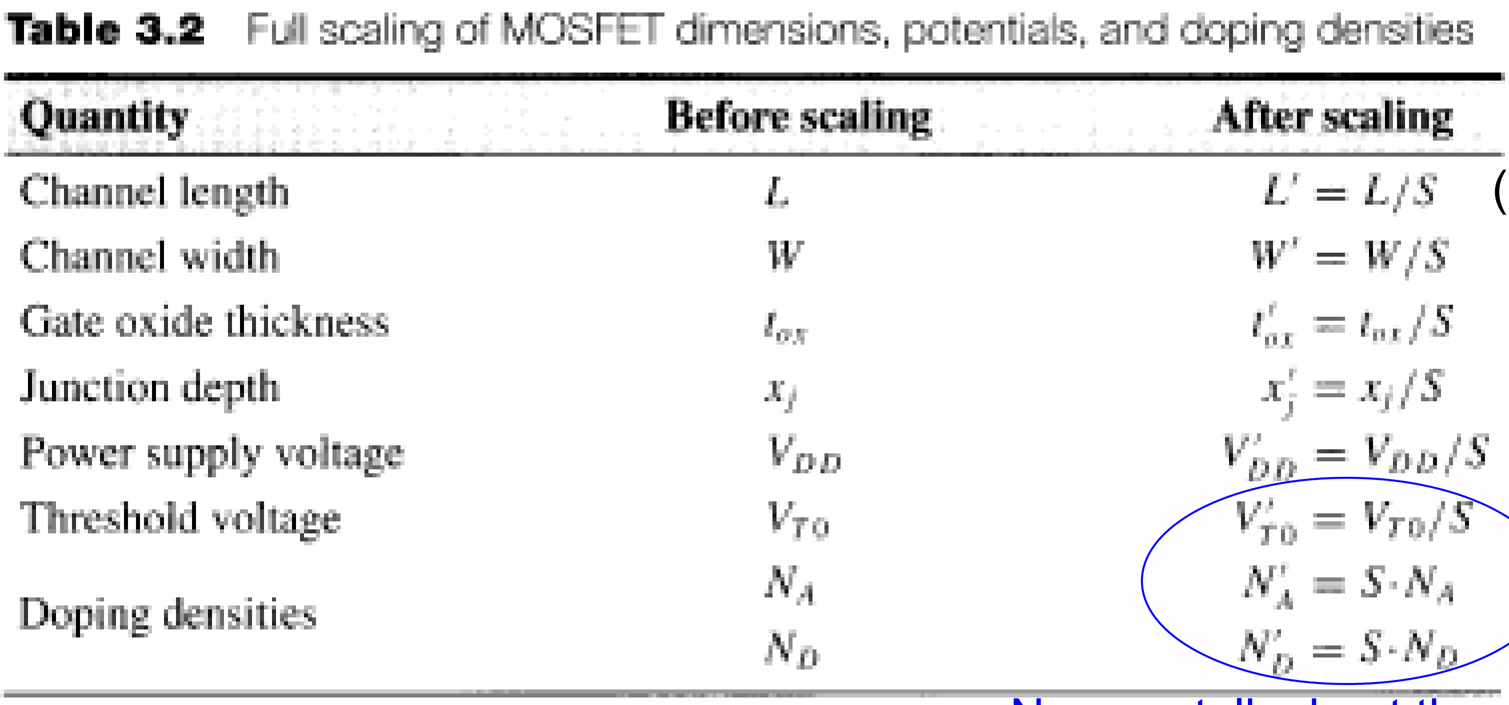


Table 3.2 Full scaling of MOSFET dimensions, potentials, and doping densities

Quantity	Before scaling	After scaling
Channel length	L	$L' = L/S$
Channel width	W	$W' = W/S$
Gate oxide thickness	t_{ox}	$t_{ox}' = t_{ox}/S$
Junction depth	x_j	$x_j' = x_j/S$
Power supply voltage	V_{DD}	$V_{DD}' = V_{DD}/S$
Threshold voltage	V_{T0}	$V_{T0}' = V_{T0}/S$
Doping densities	N_A	$N_A' = S \cdot N_A$
	N_D	$N_D' = S \cdot N_D$

($S > 1$)

Now we talk about these

$$\left.\begin{array}{r}\text{Constant filed} \boldsymbol{\rightarrow} \text{constant current } \overset{\text{(A/cm)}}{\text{density}} \\ W' = W/S\end{array}\right\} \Longrightarrow I_D' = I_D/S$$

But we did not say why. Let's now have a closer look.

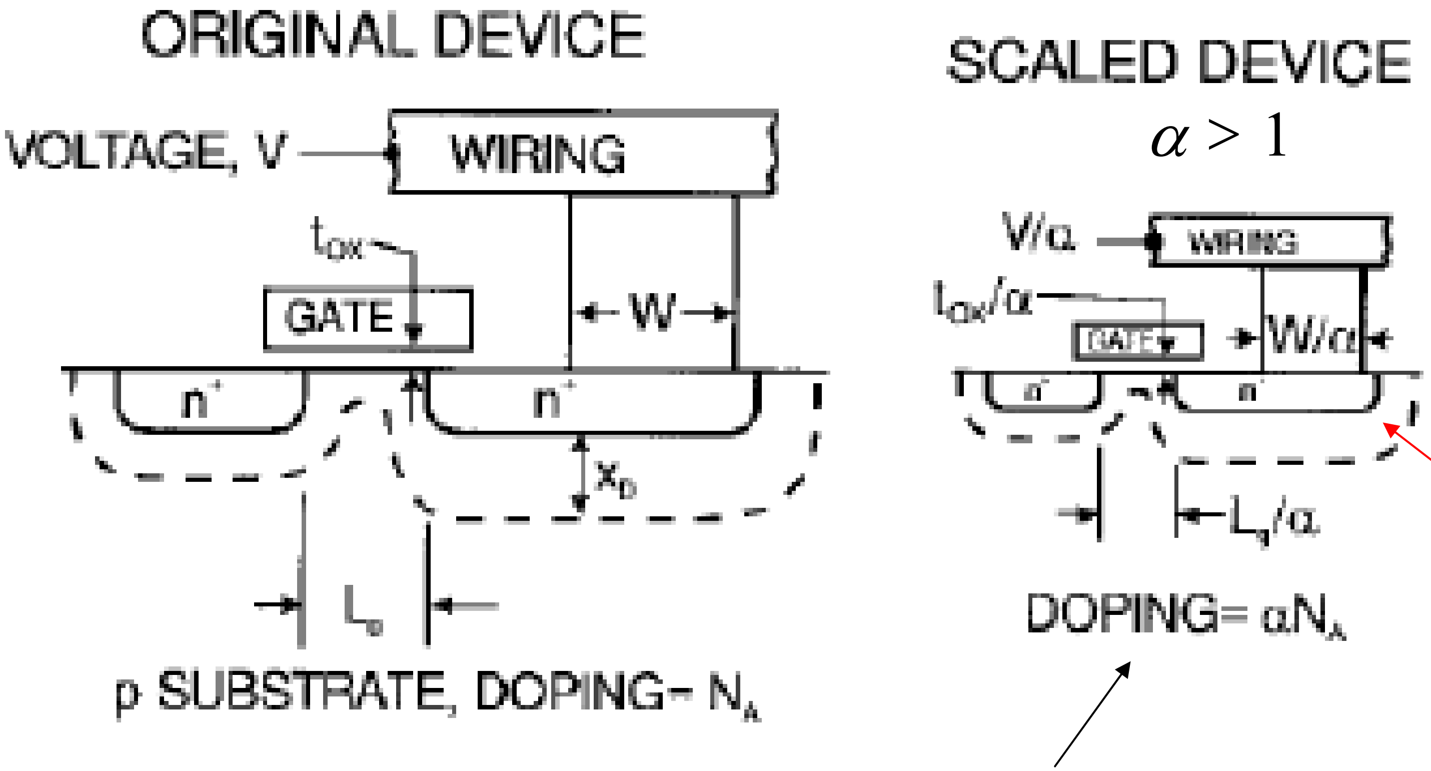
$$\left.\begin{array}{l}\text{Areal carrier charge density in channel } q\underset{\text{e's per area}}{n} = \varepsilon E_{\text{ox}} \Longrightarrow n = n' \\ \underset{\text{A/cm}}{J_s} = \mu q n E_{\text{lateral}} \\ \dfrac{\cancel{\text{cm}^2}}{\cancel{\text{V}}\text{s}}\,\dfrac{\text{C}}{\cancel{\text{cm}^2}}\,\dfrac{\cancel{\text{V}}}{\text{cm}} = \text{A/cm}\end{array}\right\} \Longrightarrow J_s' = J_s$$

$$I_D = J_s W \Longrightarrow I_D' = I_D/S$$

if we truly have constant field

$\alpha = S$

Simple constant-field scaling: $V \rightarrow \alpha V$, $I \rightarrow \alpha I$



$\alpha > 1$

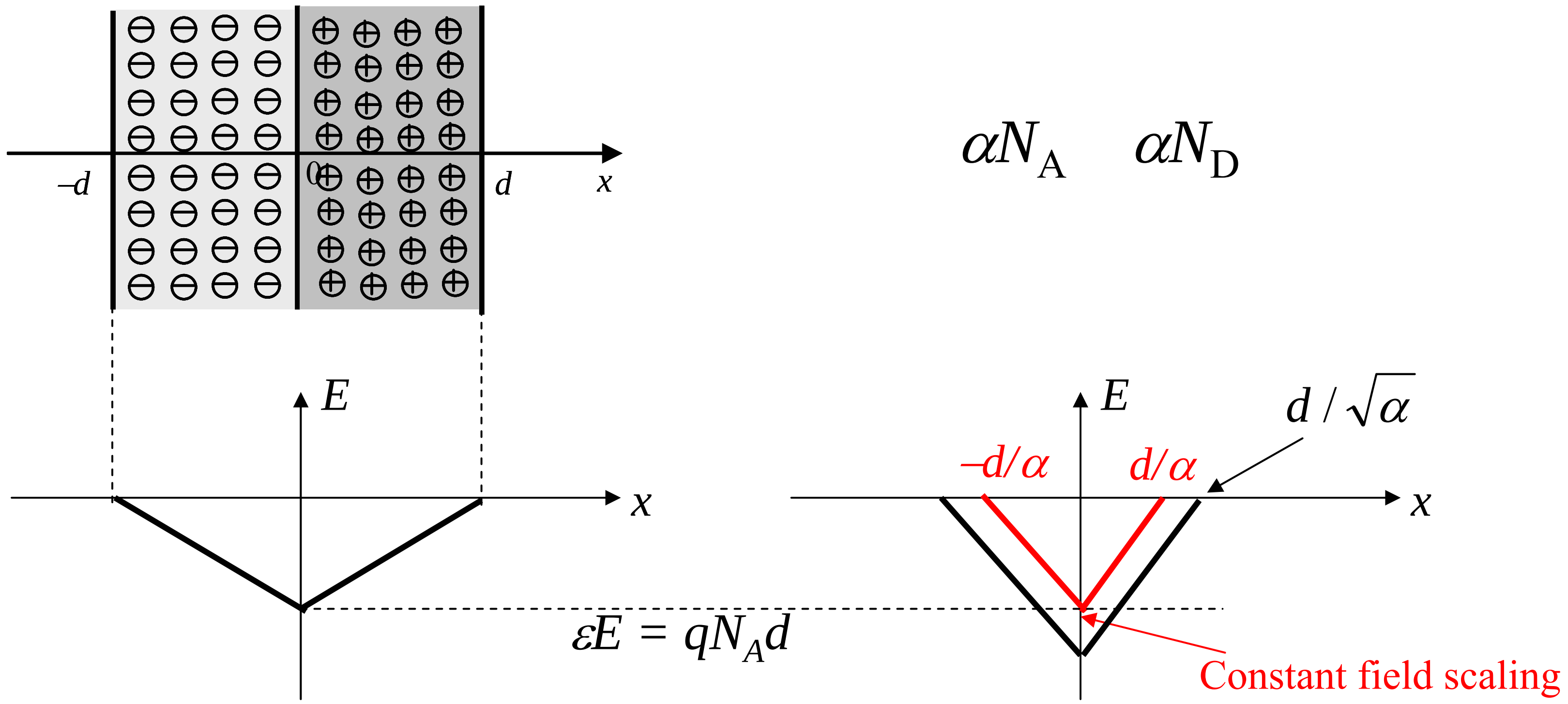
x_D/α ???
Really?

Justified?

Consider a simple pn junction within the depletion approximation (keep in mind the S/D junctions & the depletion under the channel)

Original

Scaled

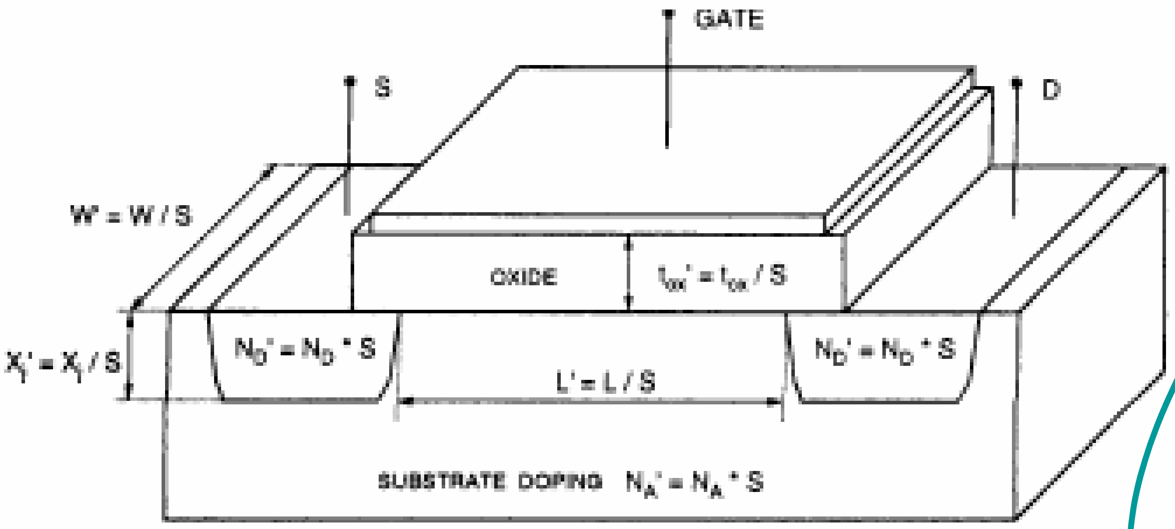


The integral of the field is the built-in potential.

But the built-in voltage is largely determined by the band gap..., not much changed in the scaled junction

Think graphically about E, potential, and the band diagram

Full Scaling (Constant-Field Scaling) (which makes (**some**) physical sense)

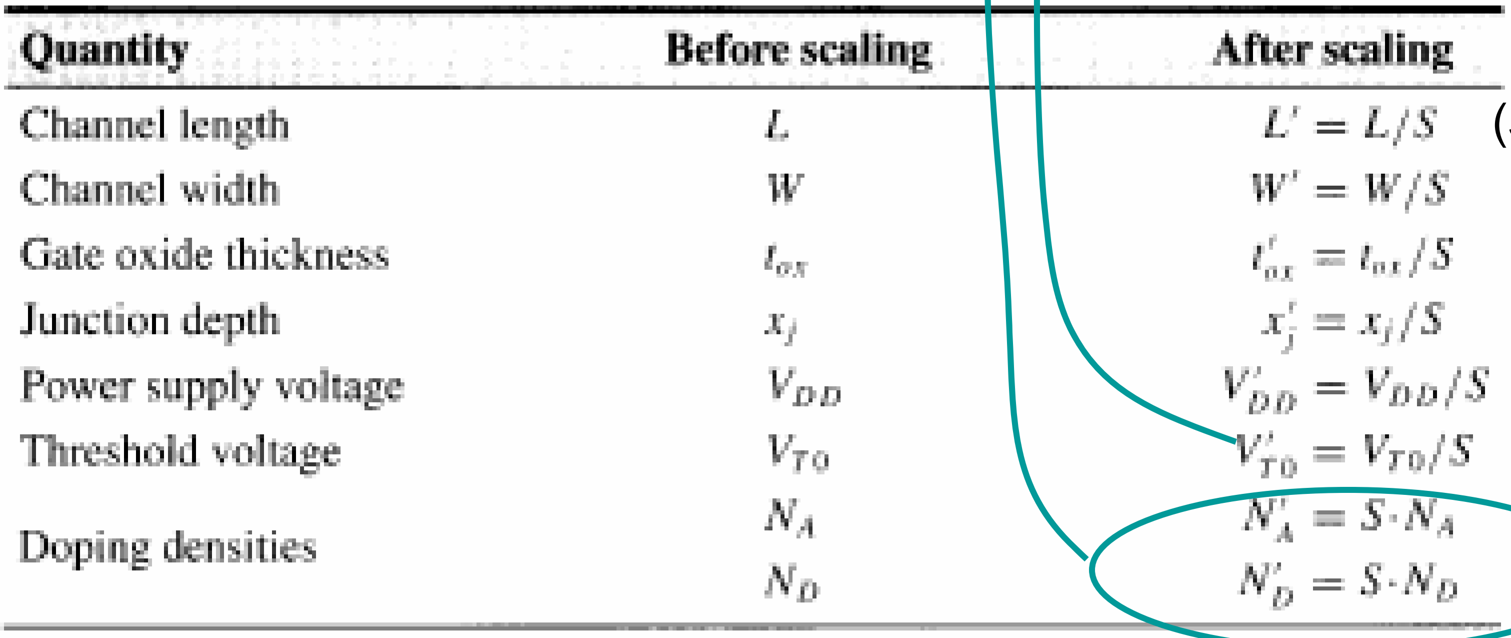


$$x_d \propto \ln(N_D N_A / n_i^2) / \sqrt{N_A}$$

$$V_{th} = \boxed{V_{FB}} + 2\phi_p + \frac{\sqrt{2\varepsilon q N_A (2\phi_p)}}{C_{ox}}$$

Doesn't scale.
Need V_{th} adjustment.

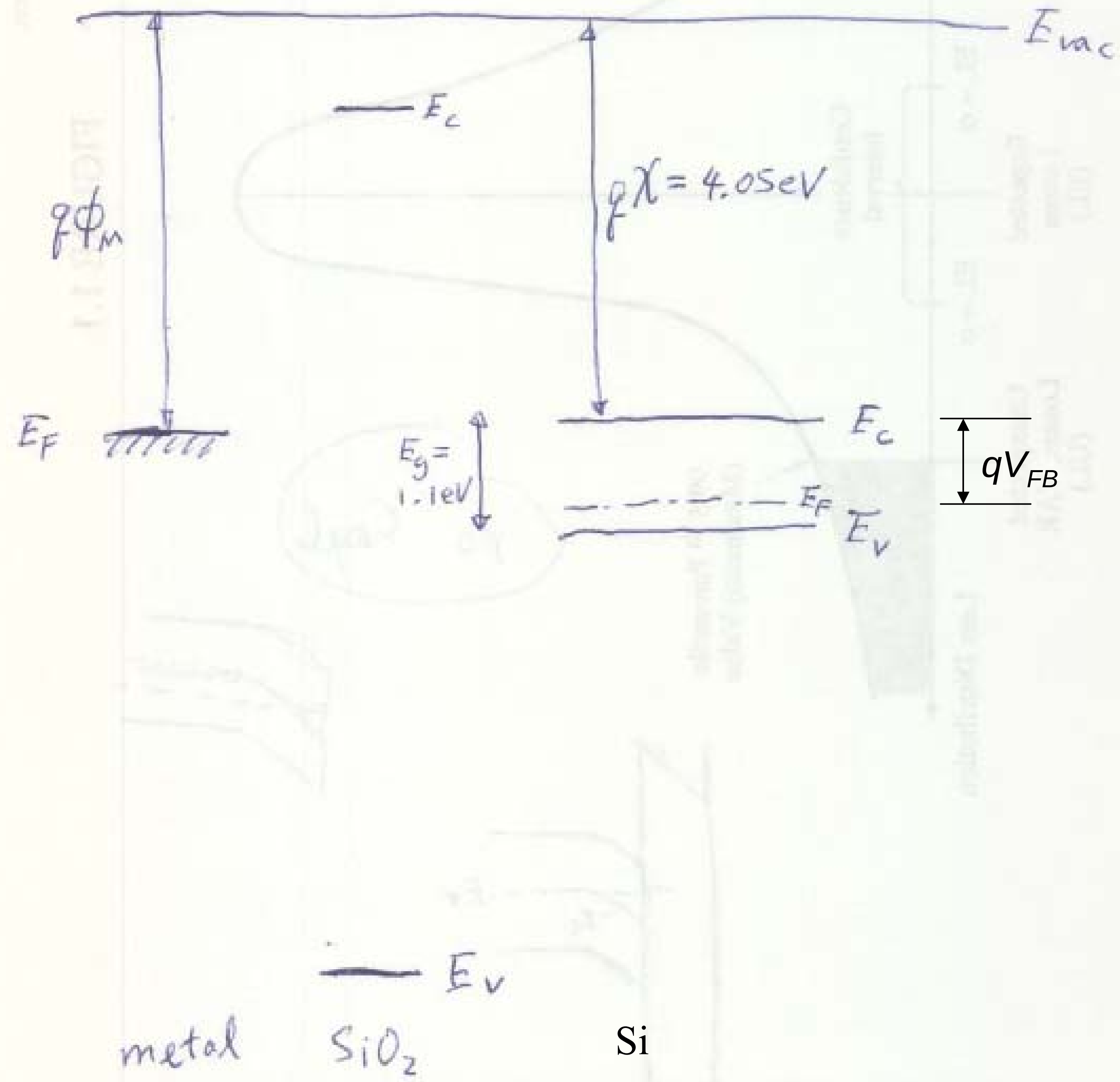
Table 3.2 Full scaling of MOSFET dimensions, potentials, and doping densities



Quantity	Before scaling	After scaling
Channel length	L	$L' = L/S$
Channel width	W	$W' = W/S$
Gate oxide thickness	t_{ox}	$t'_{ox} = t_{ox}/S$
Junction depth	x_j	$x'_j = x_j/S$
Power supply voltage	V_{DD}	$V'_{DD} = V_{DD}/S$
Threshold voltage	V_{T0}	$V'_{T0} = V_{T0}/S$
Doping densities	N_A	$N'_A = S \cdot N_A$
	N_D	$N'_D = S \cdot N_D$

(S > 1)

Not exactly

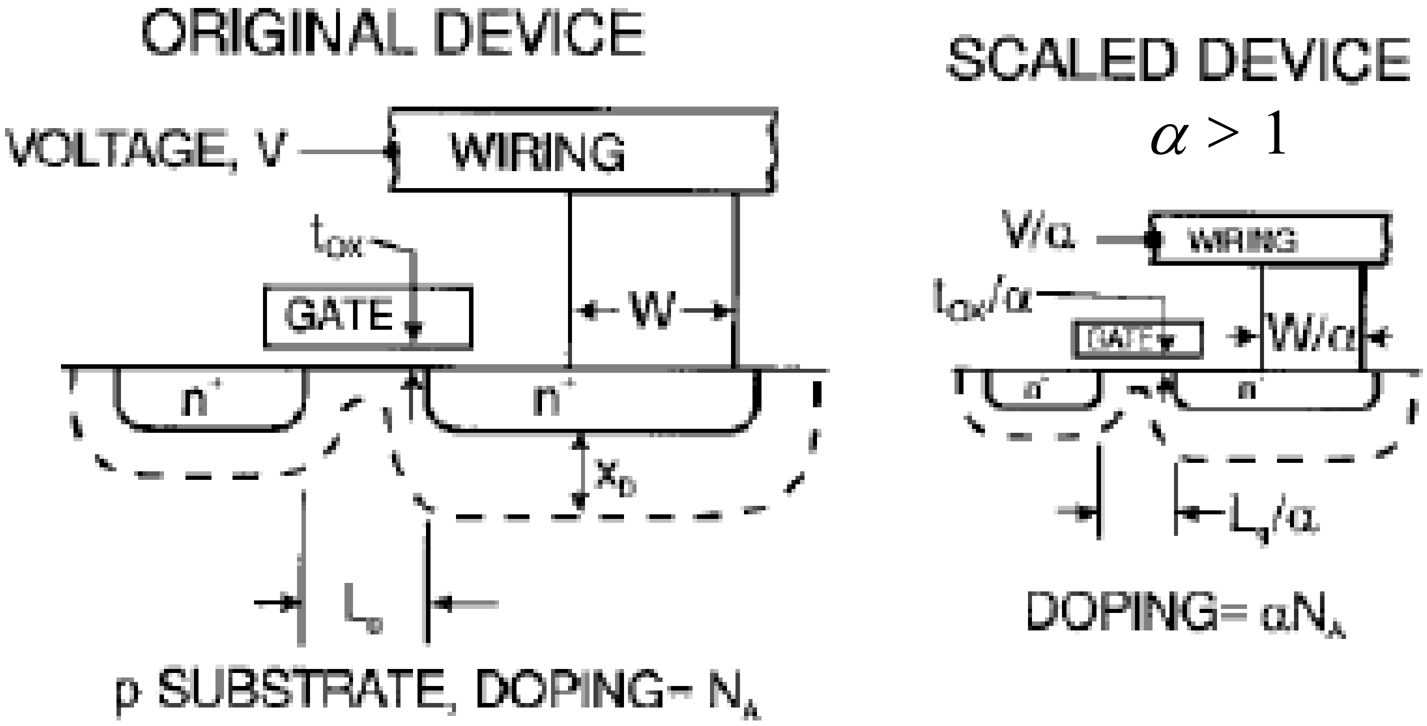


E_{vac}
E_c
$q\phi_M$
$q\chi = 4.05eV$
E_F
$E_g = 1.1eV$
E_c
qV_{FB}
E_F
E_v
E_v
metal
SiO_2
Si

Work out the details of constant-voltage scaling offline

The happy (even with the mess) scaling days are long gone.
There are many issues with scaling… Among them, electrostatic control.
(We are not going to talk about all of them.)

Simple constant-field scaling: $V \rightarrow \alpha V$, $I \rightarrow \alpha I$



But the wafer is still “infinitely thick”...

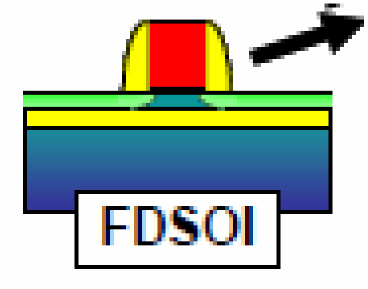
L_g	t_{ox}
800 nm	18 nm
20 nm	0.45 nm

The Si-O bond length in SiO_2 is 0.16 nm

Thicker high-k dielectric can be used, but...

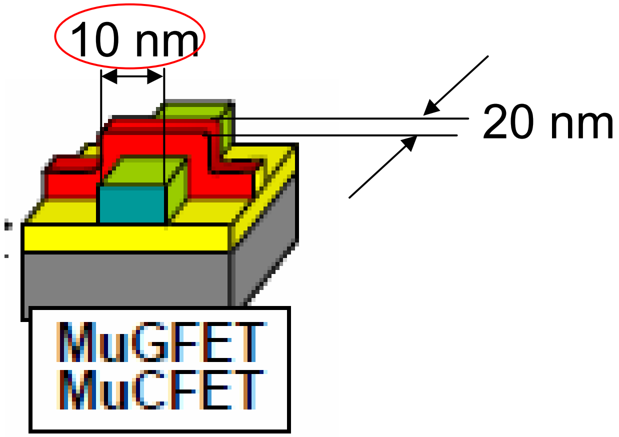
For good electrostatic control, we really need to get rid of the body (bulk)

Solutions (for now)



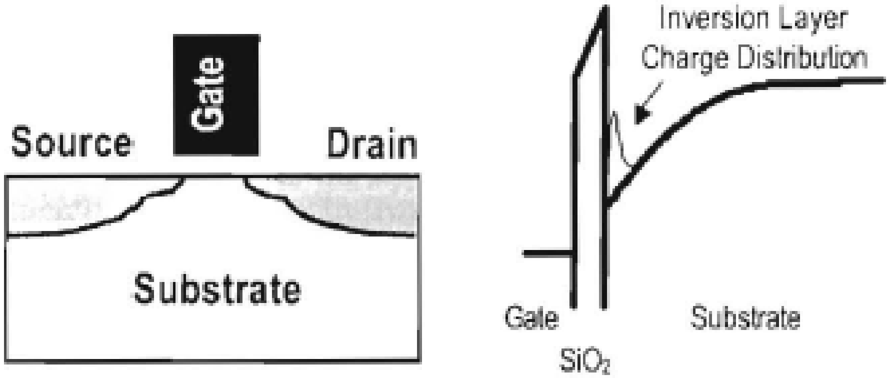
For L = 20 nm,
the Si needs to be
thinner than 5 nm.

Ultra-thin body silicon-on-insulator



FinFET, 3D FET

Further reading: “Transistor Wars,” IEEE spectrum, November 2011
http://spectrum.ieee.org/semiconductors/devices/transistor-wars

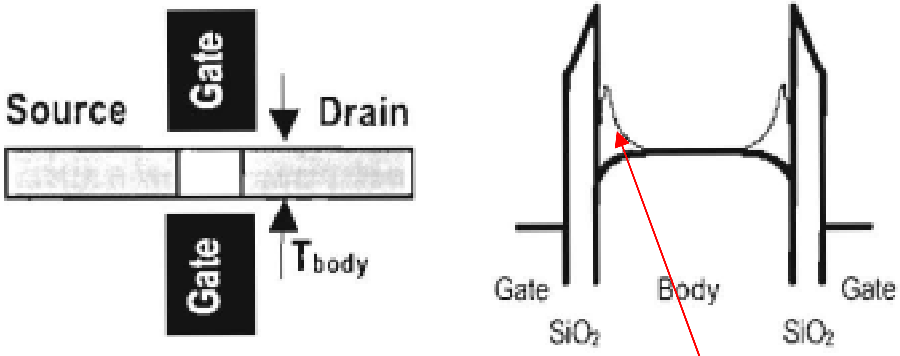


a) Bulk MOSFET

Deplete and then invert.

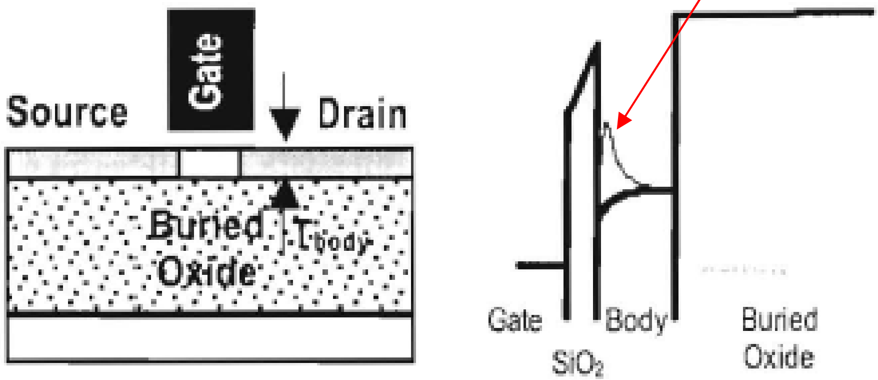
The bulk is a path for leakage.

Will need very high N_A for the substrate, to scale down the x_d of the S/D junctions, as well as the x_{dmax} of the FET channel. This will cause high S-to-D leakage.



b) Double-Gate MOSFET

Carrier density



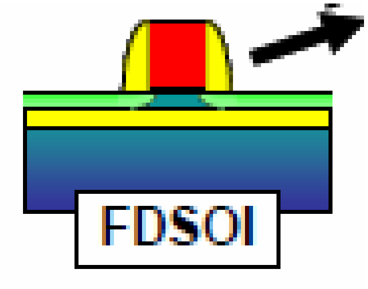
c) Ultra-Thin Body MOSFET

Thin bodies don't have to be doped.
➔ No need for depletion
➔ Lower vertical fields

Why is SOI a challenge?
http://en.wikipedia.org/wiki/Silicon_on_Insulator

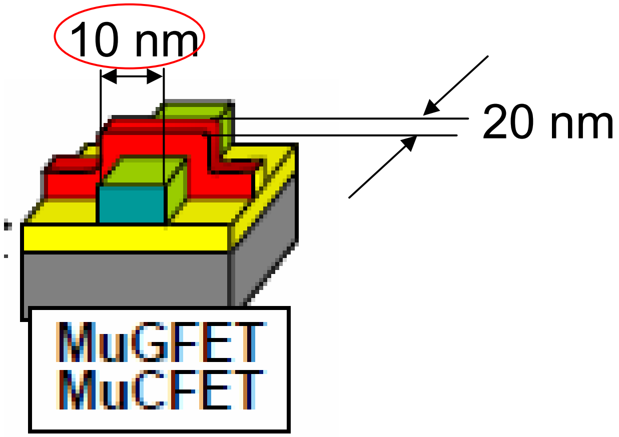
For good electrostatic control, we really need to get rid of the body (bulk)

Solutions (for now)



For L = 20 nm,
the Si needs to be
thinner than 5 nm.

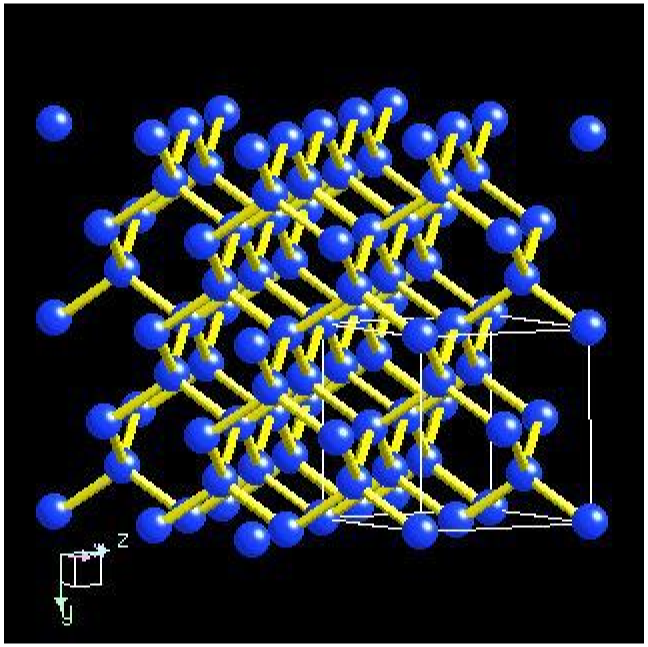
Ultra-thin body silicon-on-insulator

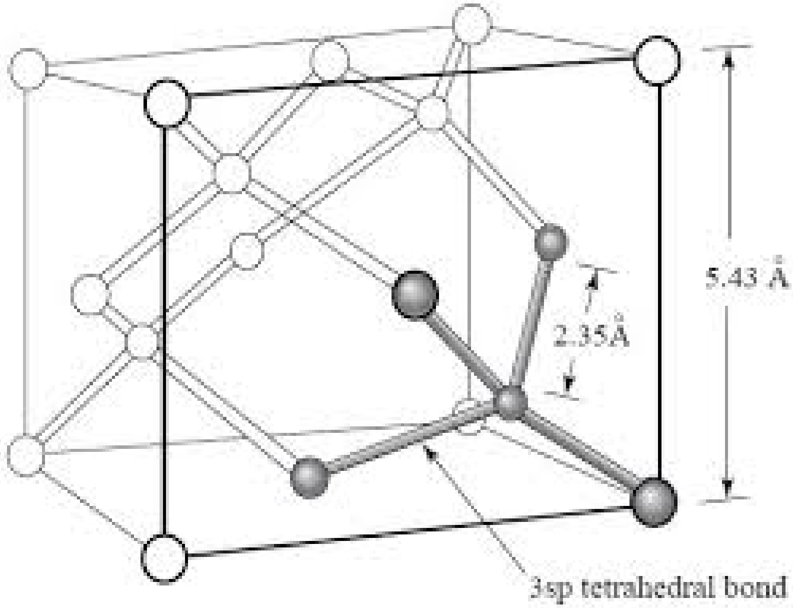


FinFET, 3D FET

Besides technical challenges, any issues with making the fin thinner?

Crystal structure of Si

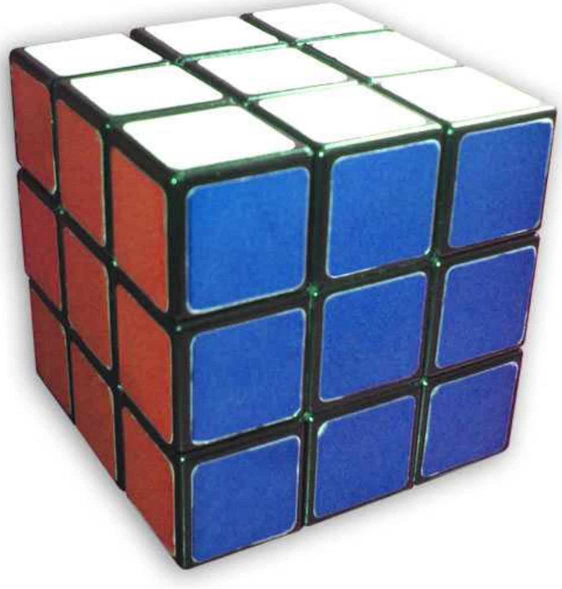




http://www.webelements.com/silicon/crystal_structure.html

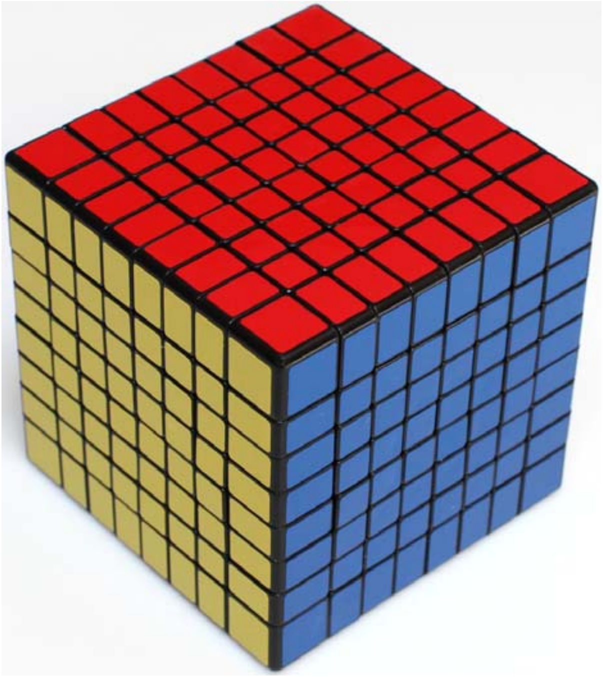
http://onlineheavytheory.net/silicon.html

If several monolayers thin, is Si still the Si as we know it?



$3^3 = 27$

$(3–2)^3 = 1$



$10^3 = 1000$

$(10–2)^3 = 512$

Half of the small cubes are on the surface!

Can Moore's Law Go On Forever?

- People have had doubts for many years
- But so far so good (?)
- In the past, the doubts were more about processing limits than physics limits
- Processing limits have been overcome
  And probably will continue to be overcome
- Are we hitting the physics limits?

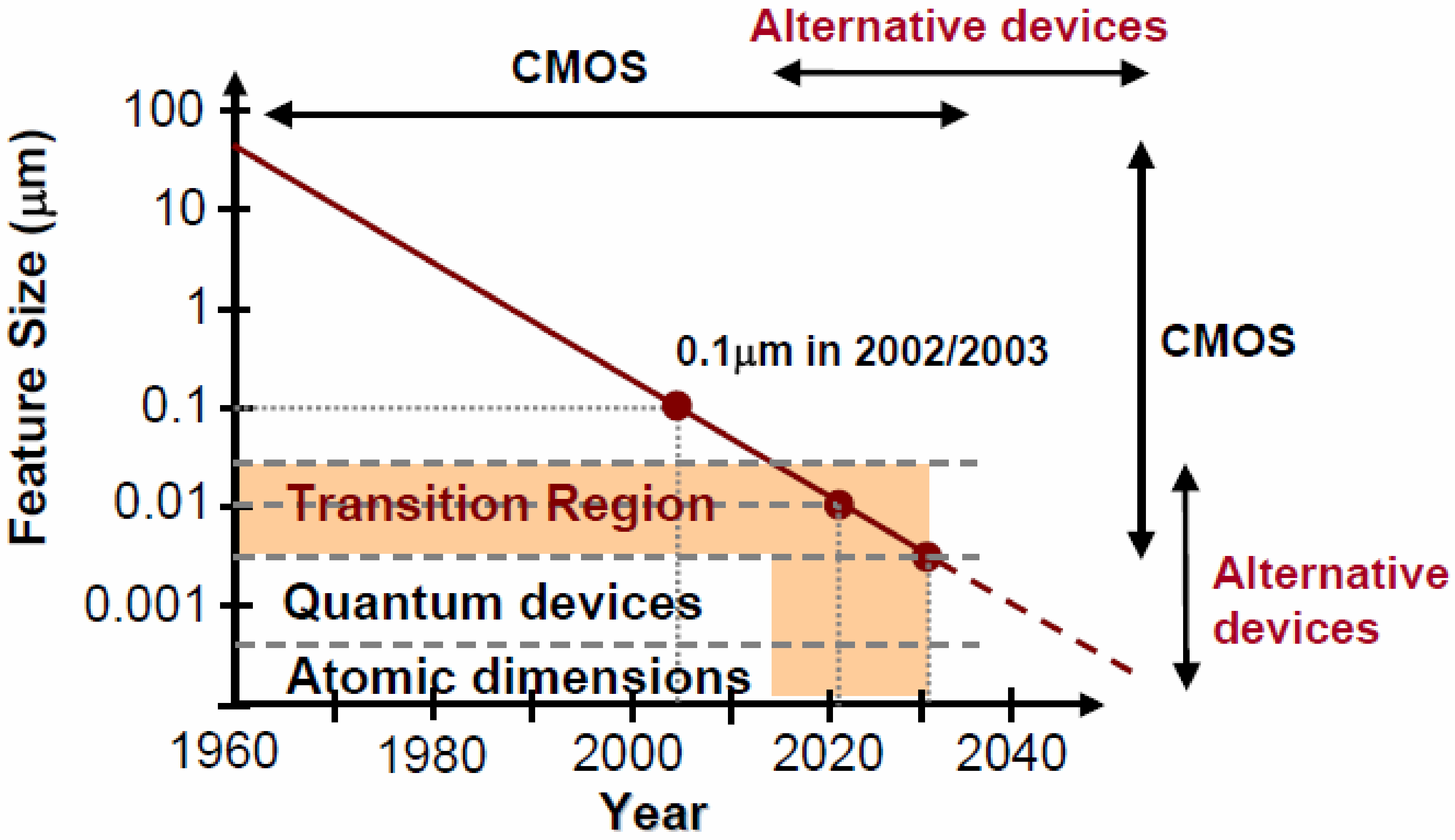
Forget about details.
The crystal constant of Si is 0.54 nm.

Fundamental scaling limit: 10 nm wall?
10 nm is only ~19 crystal constants!

Do the (semi-classical) semiconductor physics theories we rely on still work?

Are there alternatives to scaling to achieve faster devices?

Nano-reality: CMOS IC evolution

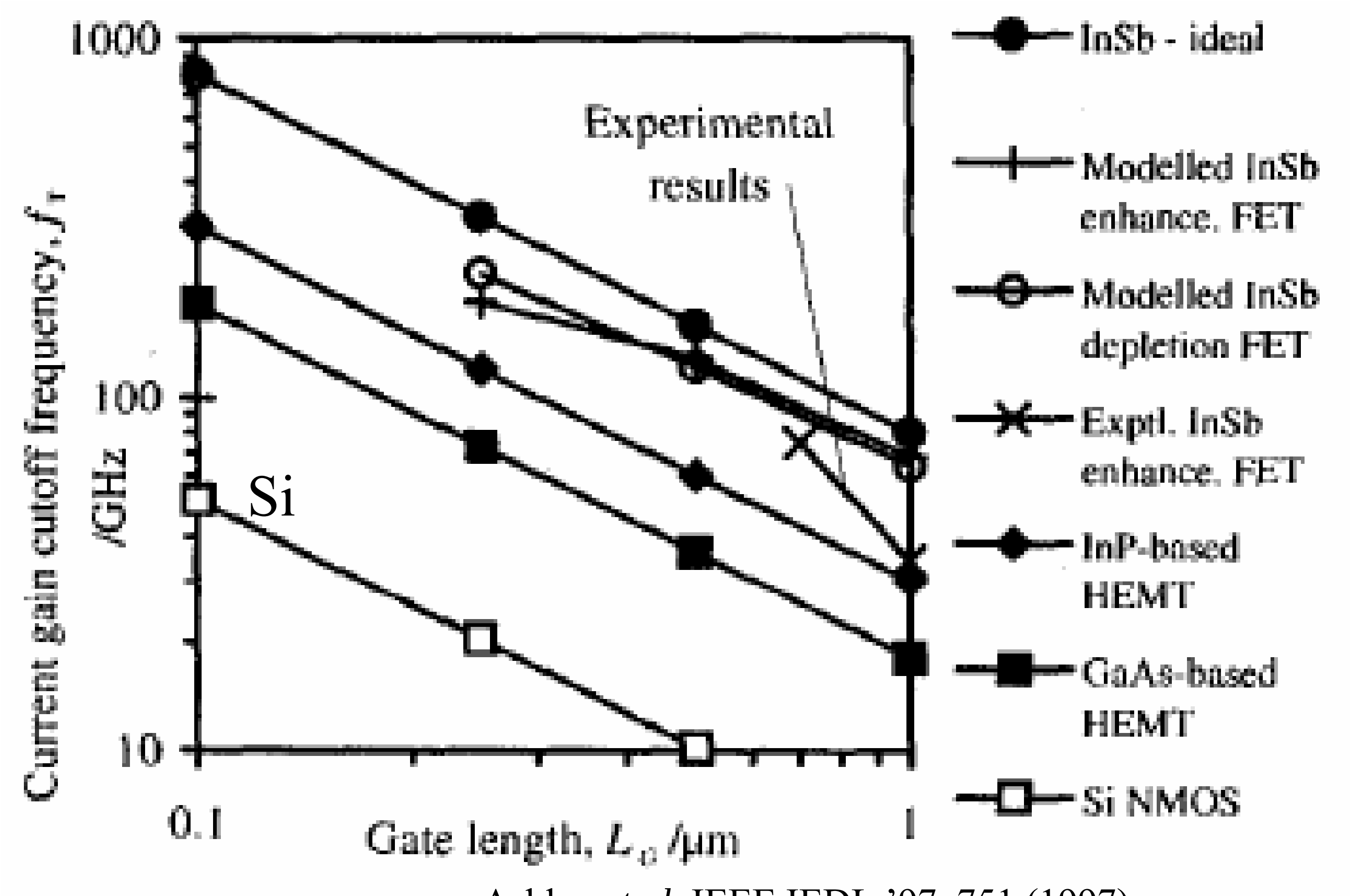


Adapted from J.D. Plummer, Proceedings of IEEE, 2001.

One alternative is semiconductors in which electrons travel faster.
(If our goal were only to make the device faster.)

But economics is in the driver’s seat…

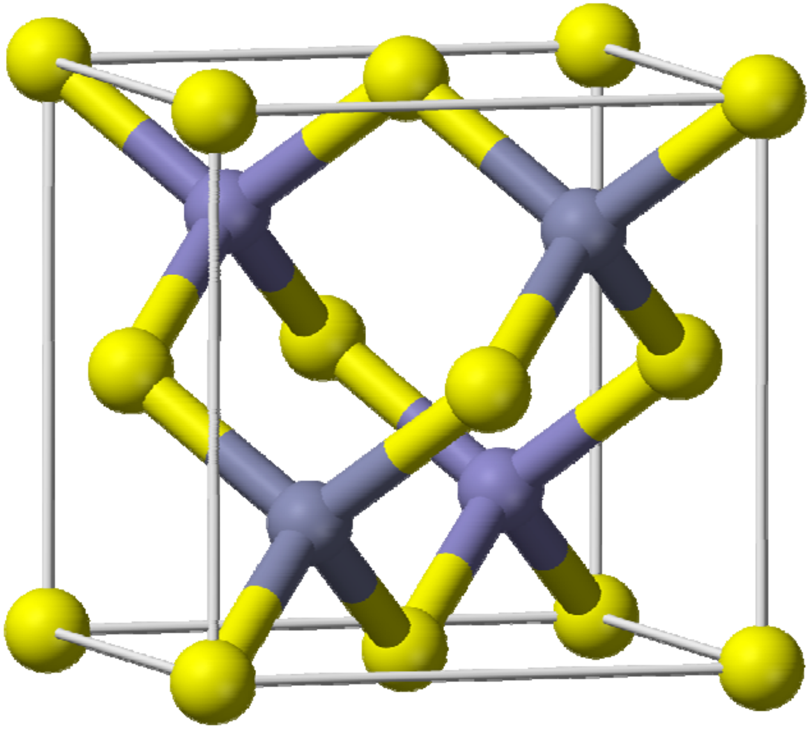
$f_T = v_{sat}/(2\pi L)$



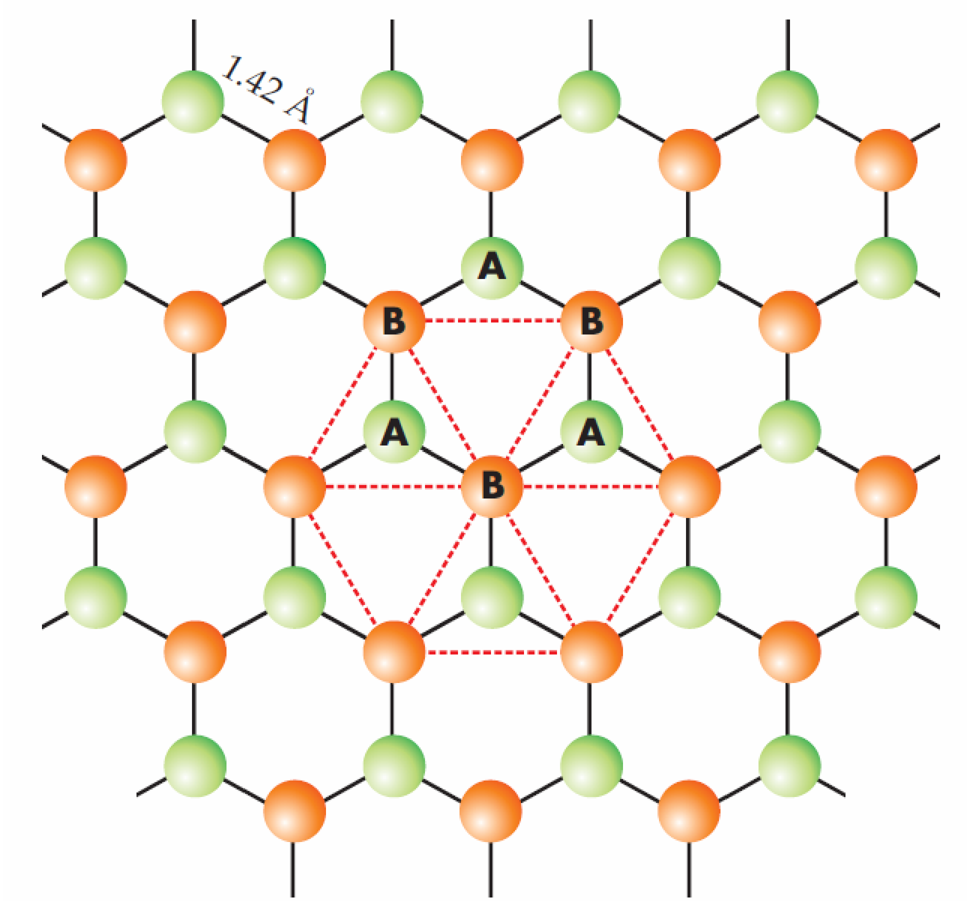
Ashley *et al*, IEEE IEDL ’97, 751 (1997).

Reading: Ye, III-V MOSFETs (posted on course website).

How thin can Si (or any 3D stuff) go?
For a 3D material, if we make it a few atoms thin, it's no longer that material as we know it!
(Recall that 10X10X10 cube)



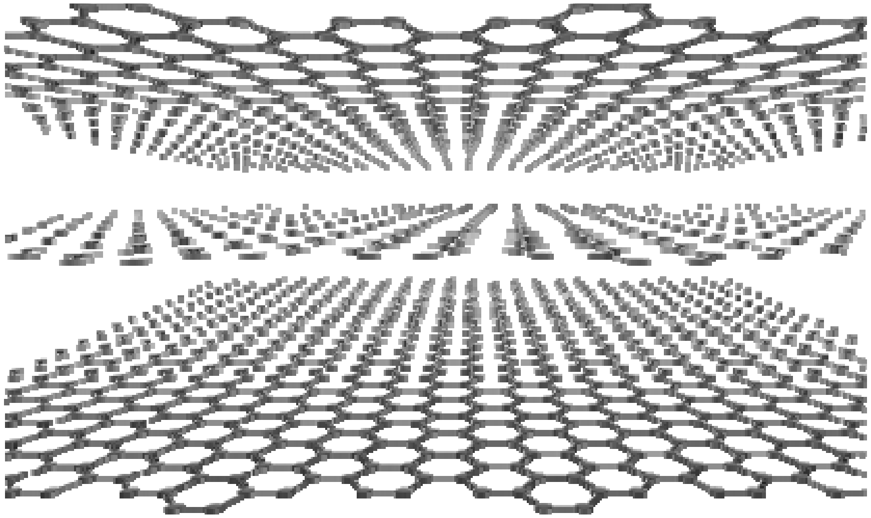
The need for lower dimensional materials



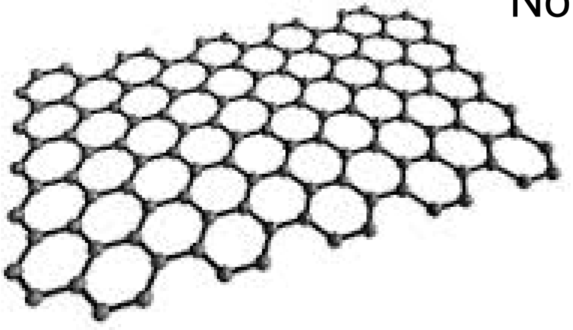
Graphene is 2D.
It's 1-atom thin by nature.
The "ultimate SOI" if we put it on an insulator.

If it ever (?) becomes the post-Si semiconductor, its 2D nature (actually, we may need to turn it into 1D) probably deserves more kudos than its fast moving electrons.

At the end of the road, we look for alternatives.
We want things that are inherently low dimensional (2D or 1D)

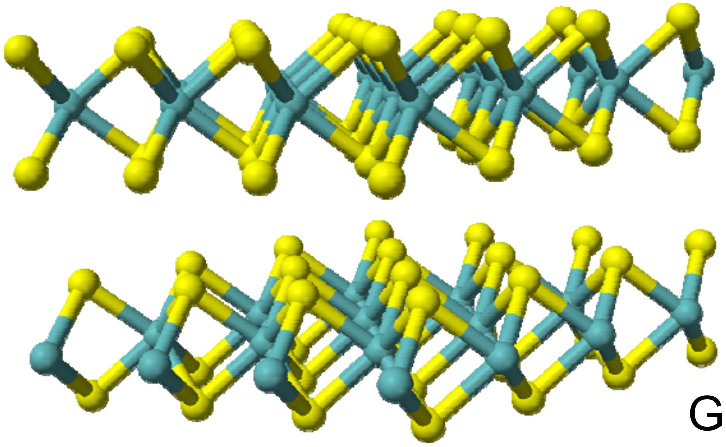


Graphite: 3D but layered (w/ van der Waals gaps)



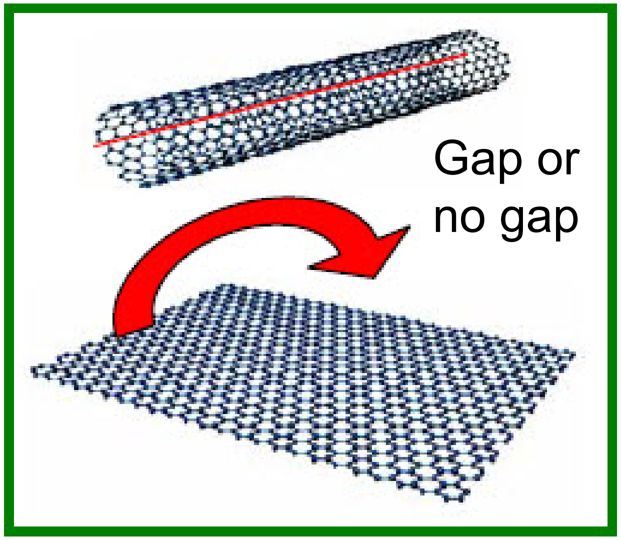
No gap

Graphene: 2D



Gap

MoS_2

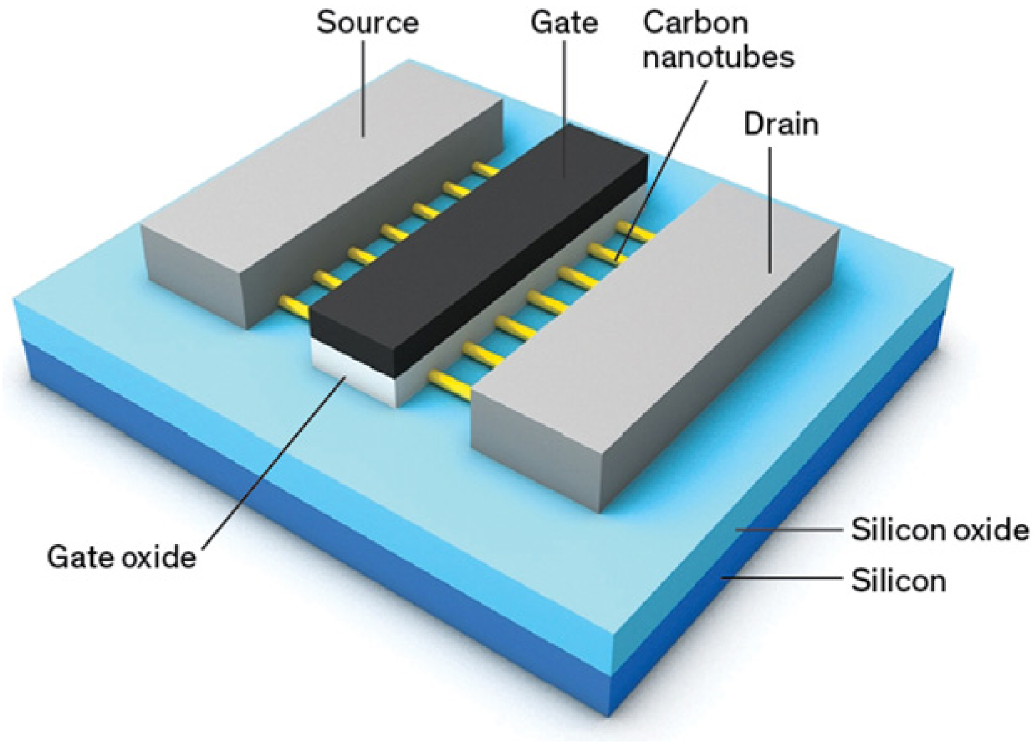


Carbon nanotube: (quasi-)1D

Carbon nanotube FETs are considered promising.

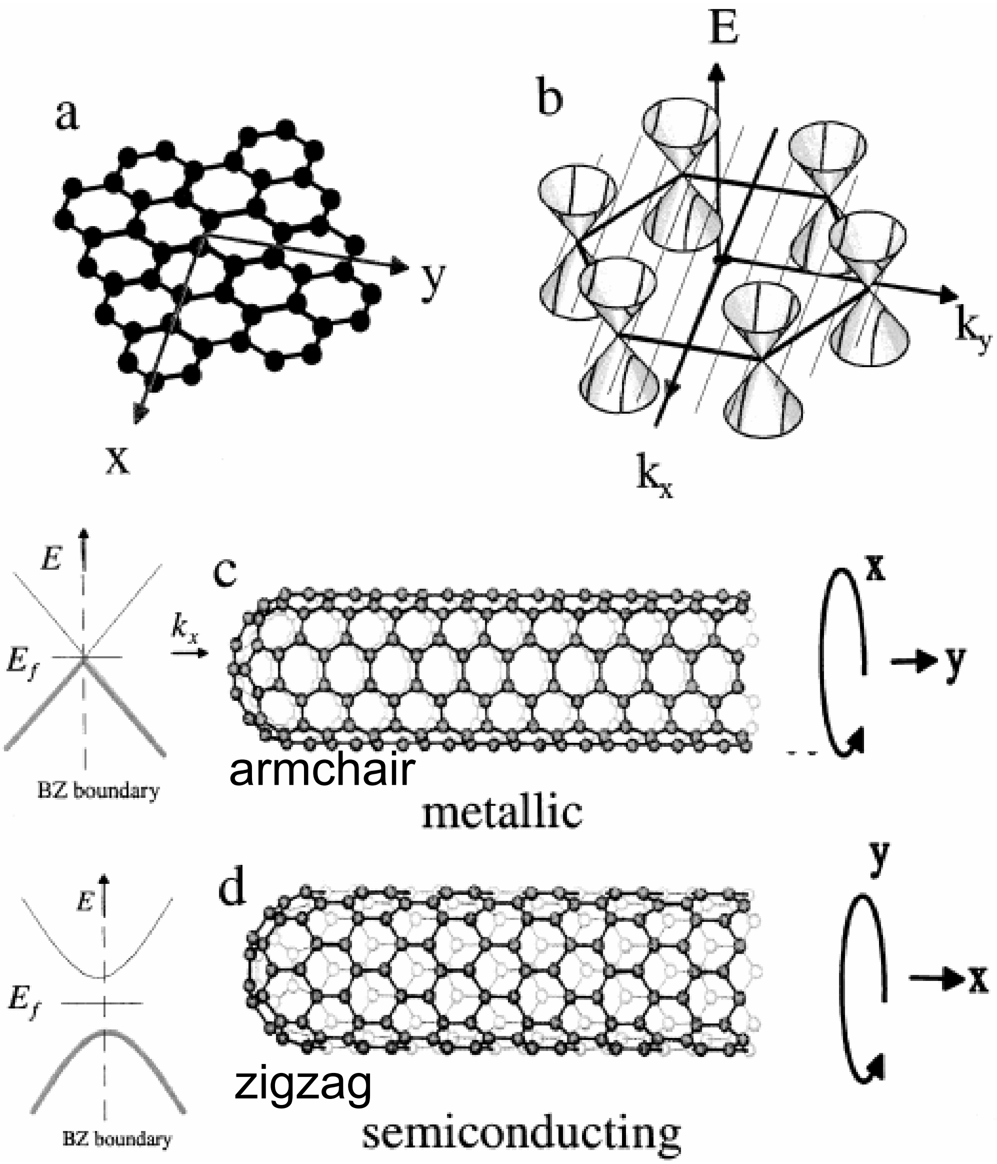
1D. Good electrostatic control.

Challenges:
To achieve uniform diameter and chirality (for uniform band gap)
To place them into a dense, parallel array



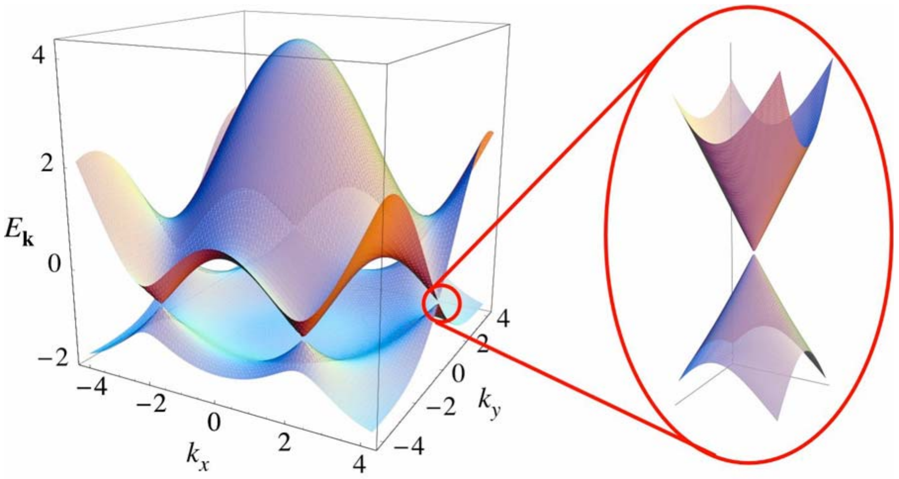
https://spectrum.ieee.org/semiconductors/devices/how-well-put-a-carbon-nanotube-computer-in-your-hand

Band structures of single-wall carbon nanotubes



McEuen et al, IEEE Trans on Nanotech 99, 78 (2002)

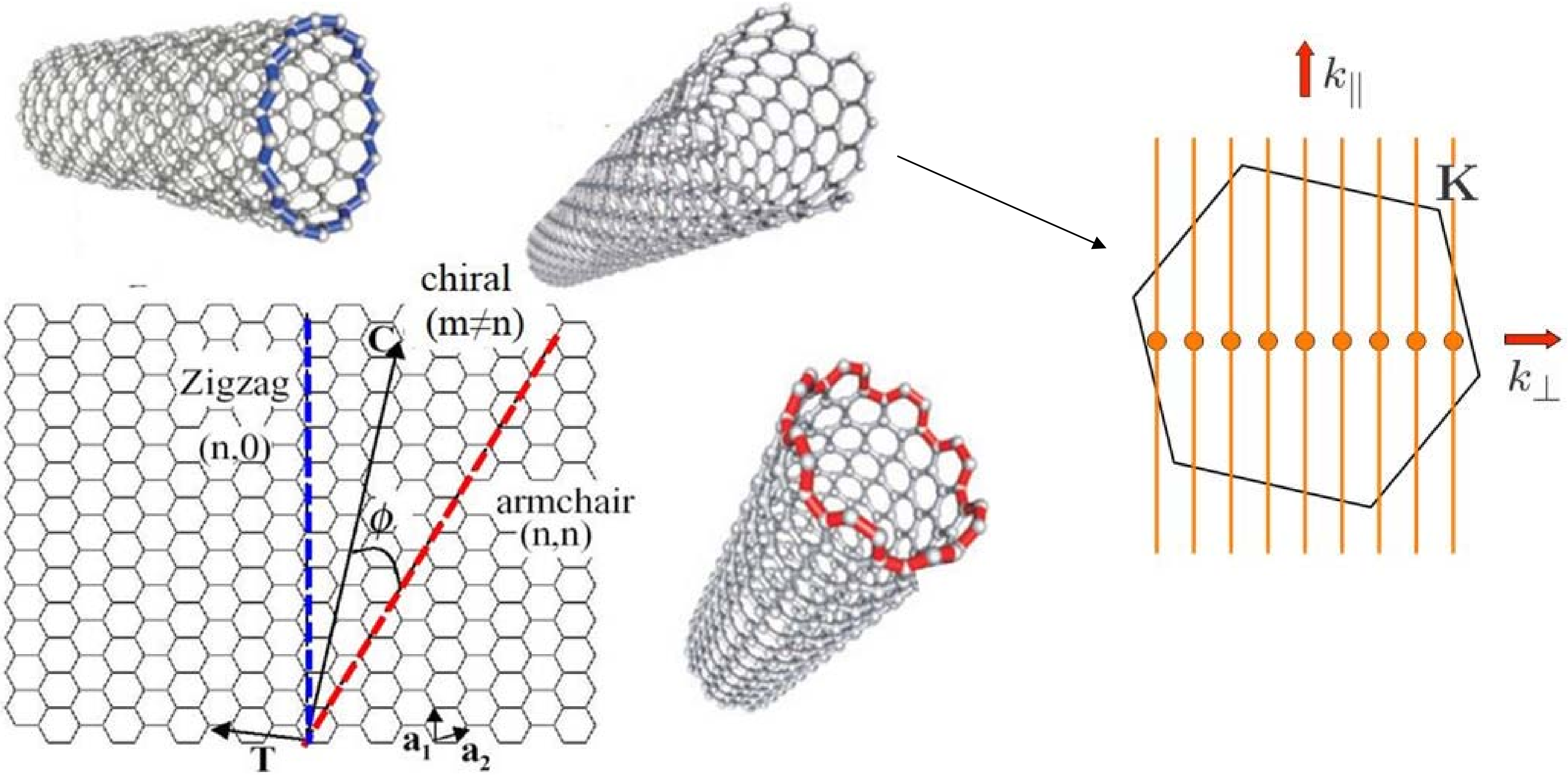
Band structure of graphene



http://exciting-code.org/carbon-graphene-from-the-ground-state-to-excitations

Simple theory to construct nanotube band structure:
Quantizing **k**

Carbon nanotube indexing

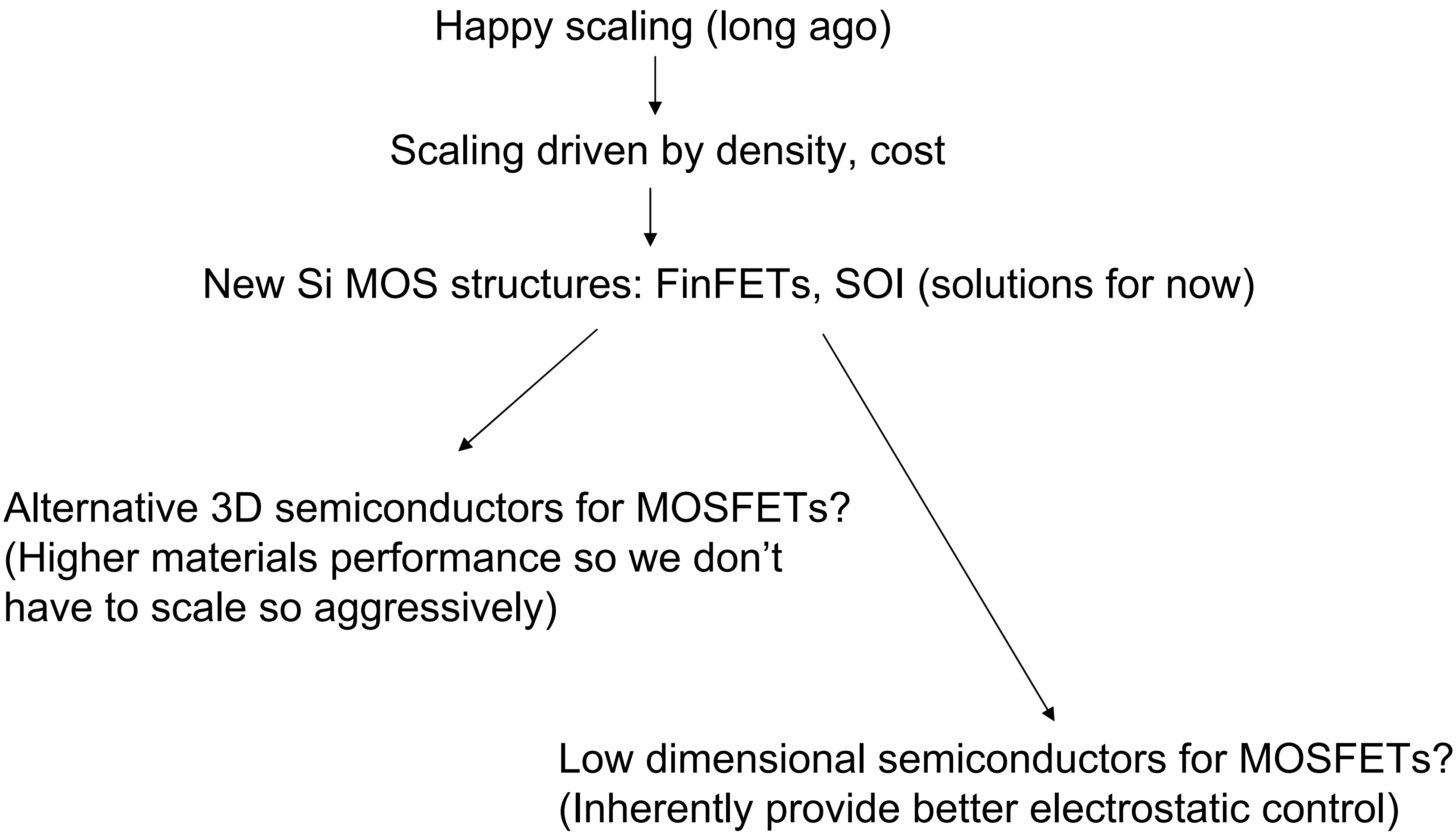


https://www.frontiersin.org/articles/10.3389/fchem.2015.00059/full

Simple theory (not exact):

$n - m$ = multiple of 3 ➔ metallic
Therefore, armchair (n,n) always metallic,
zigzag (n,0) metallic when n = multiple of 3,
semiconducting otherwise.

Now you see how challenge it is to get all tubes to have the same bandwidth



Happy scaling (long ago)

Scaling driven by density, cost

New Si MOS structures: FinFETs, SOI (solutions for now)

Alternative 3D semiconductors for MOSFETs?
(Higher materials performance so we don’t
have to scale so aggressively)

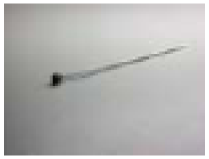
Low dimensional semiconductors for MOSFETs?
(Inherently provide better electrostatic control)

Alternative devices, architectures, systems???

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

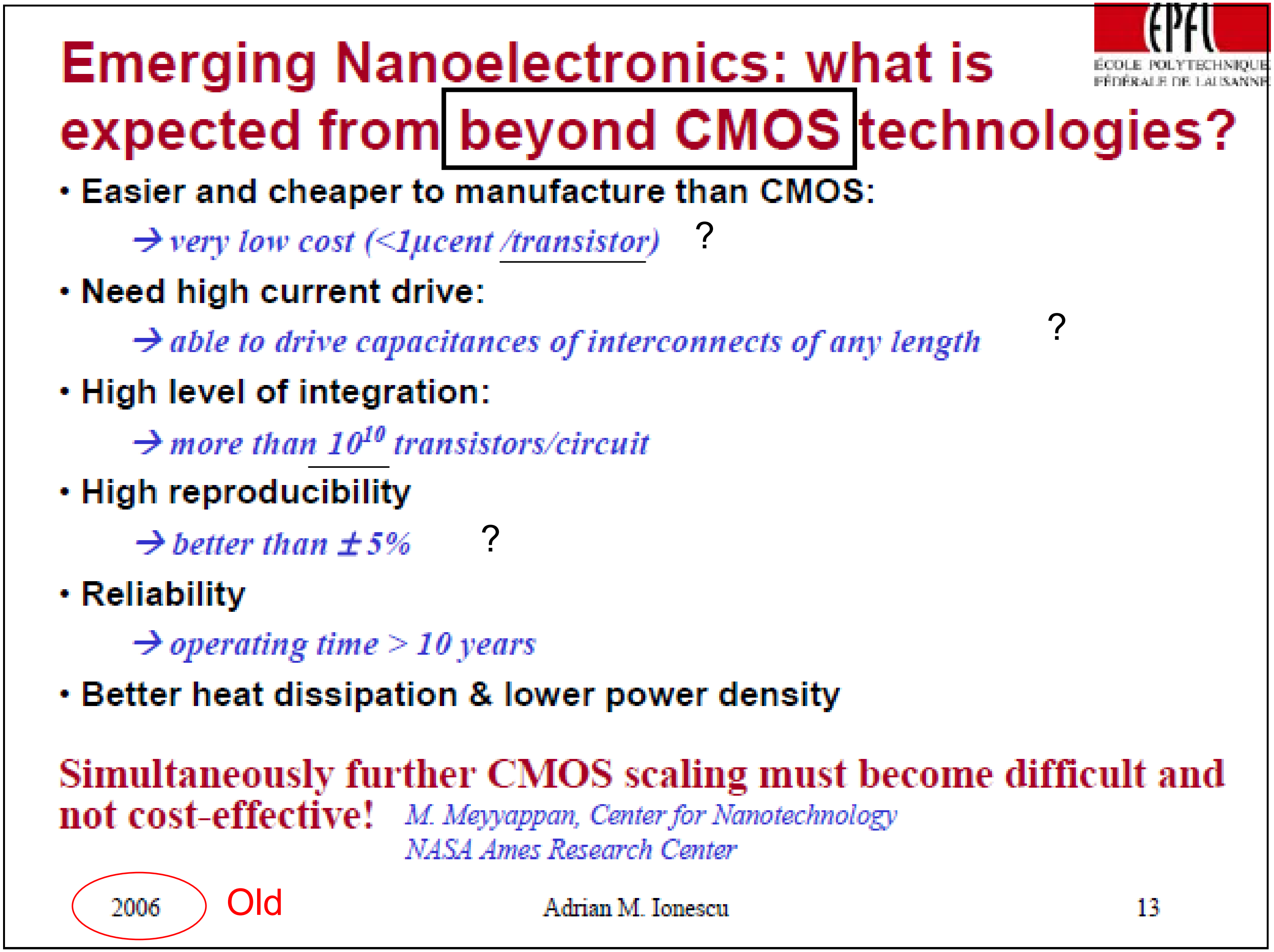
Invitation To Enter the NANO-World: Information on a small scale

There's Plenty of Room at the Bottom
An Invitation to Enter a New Field of Physics
by Richard P. Feynman
December 29th, 1959



Why cannot we write the entire 24 volumes of the Encyclopedia Brittanica on the head of a pin?

A future filled with tiny, molecule-sized computers-fast and powerful enough to do things like translate conversations on the fly or calculate complex climate models-may be closer than people think…
American Association for the Advancement of Science (AAAS), Annual Meeting (Boston, February, 2002)



Emerging Nanoelectronics: what is expected from beyond CMOS technologies?

- Easier and cheaper to manufacture than CMOS:
  - → *very low cost (<1µcent /transistor)* ?
- Need high current drive:
  - → *able to drive capacitances of interconnects of any length* ?
- High level of integration:
  - → *more than 10^{10} transistors/circuit*
- High reproducibility
  - → *better than ± 5%* ?
- Reliability
  - → *operating time > 10 years*
- Better heat dissipation & lower power density

Simultaneously further CMOS scaling must become difficult and not cost-effective! *M. Meyyappan, Center for Nanotechnology*
NASA Ames Research Center

2006 Old