

The Future of CMOS

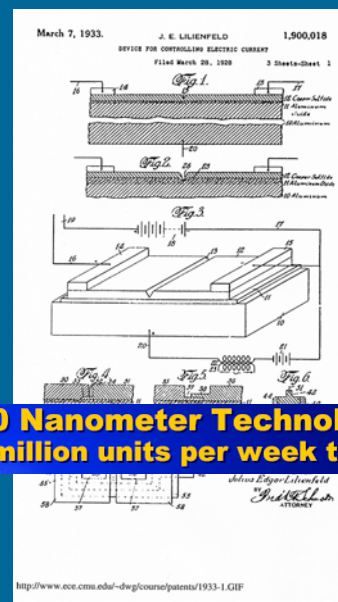
David Pulfrey



1

CHRONOLOGY of the FET

- 1933 Lilienfeld's patent (BG FET)
- 1965 Commercialization (Fairchild)
- 1991 "The most abundant object made by mankind" (C.T. Sah)
- 2003 The 10 nm FET (Intel)



**90 Nanometer Technology
 (1 million units per week today)**

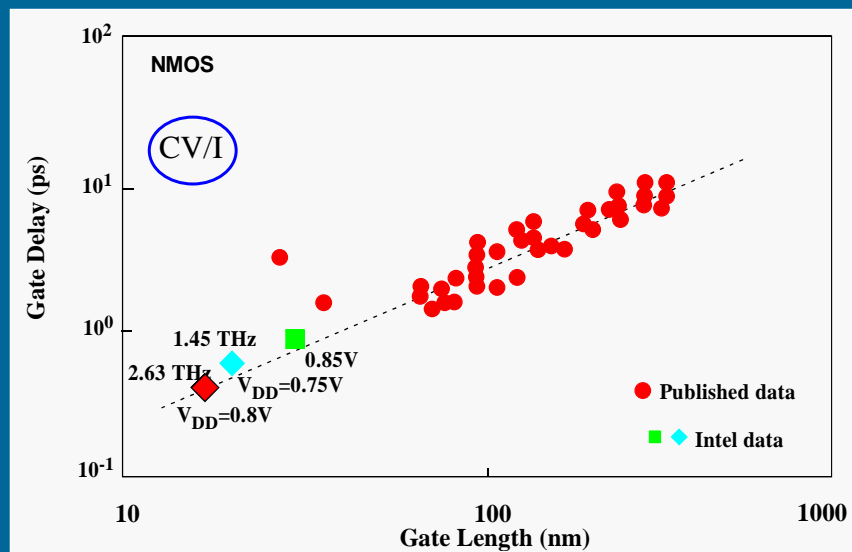
2

SCALING: WHY DO IT?

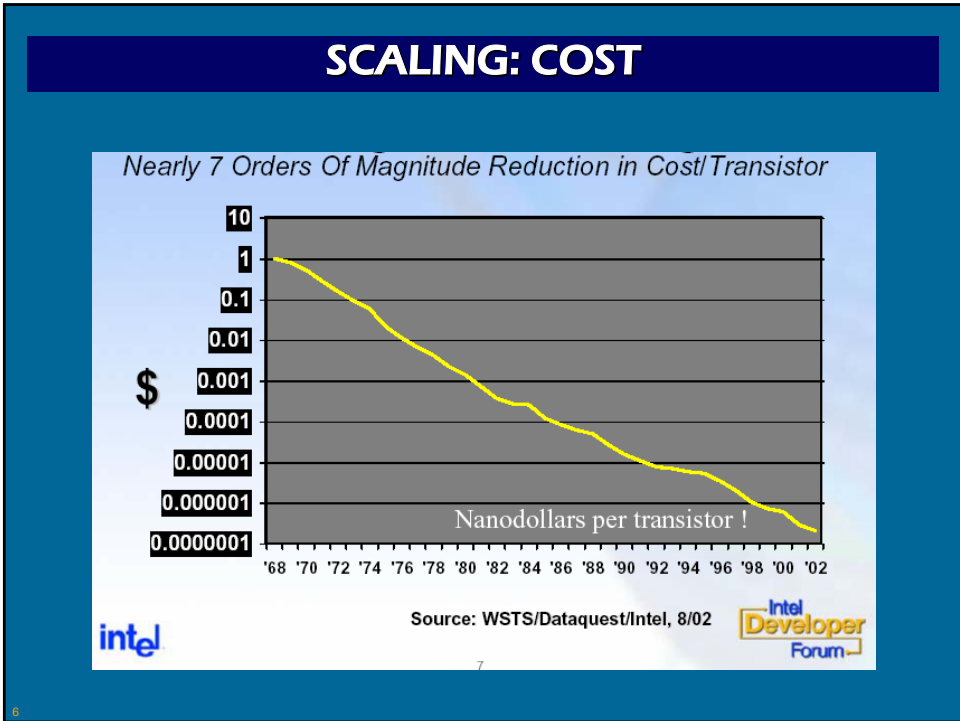
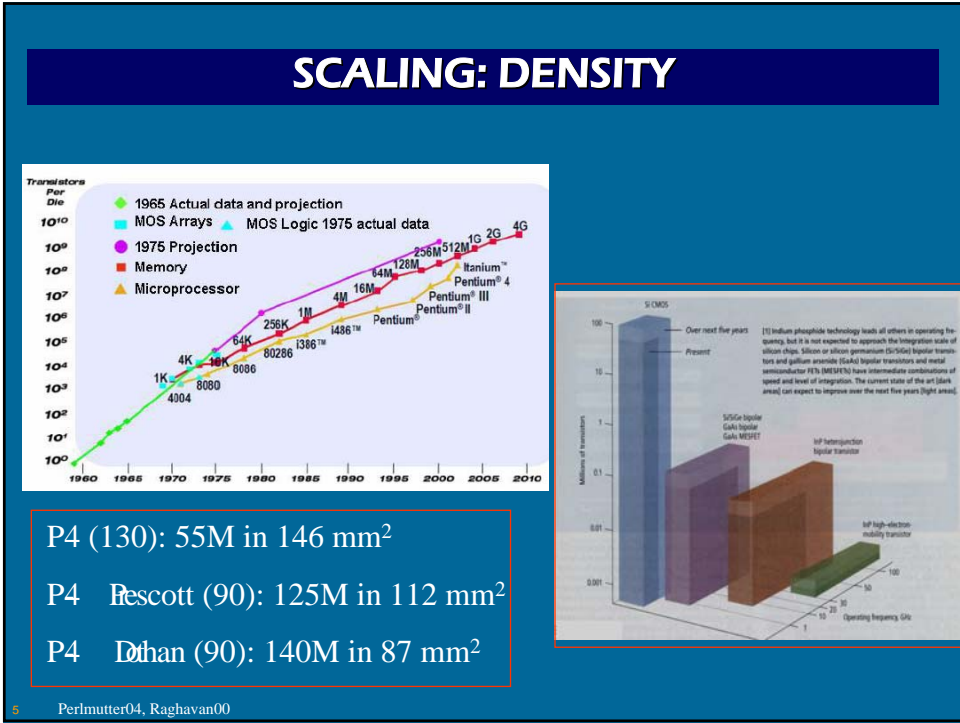
- Increase speed
- Increase density
- Reduce cost (?)

3

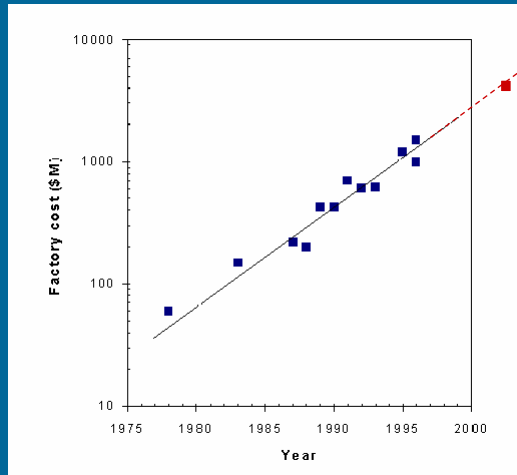
SCALING: SPEED



Morkoc04



SCALING: FACTORY COST

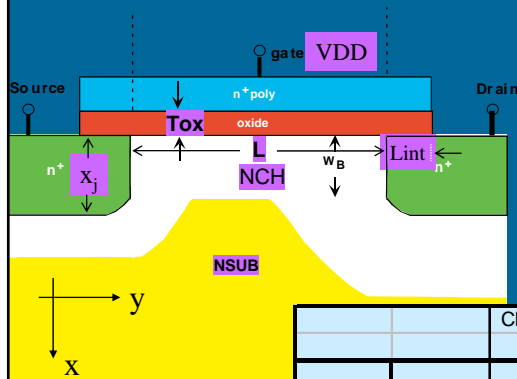


Intel 300mm

Moore's Law
for Fabs!

7 MIT02

THE SHRINKING FET

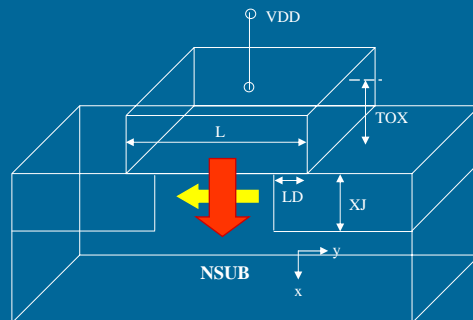


- L_{eff} reduced 30X
- But devices are still "well tempered"

		CMOS 3 1987	CMOS P18 2001	CMOS P13 2002	90NM 2003
L	nm	3000	180	130	100
LINT	nm	700	10	0	2.5
XJ	nm	1000	160	190	150
TOX	nm	85	4.1	2.8	2.3
NCH	cm ⁻³	1.00E+16	3.90E+17	6.15E+17	8.37E+17
VDD	V	5.0	1.8	1.2	1.0
VTHO	V	0.95	0.47	0.35	0.24

8

THE SIGNIFICANCE OF E_x AND E_y



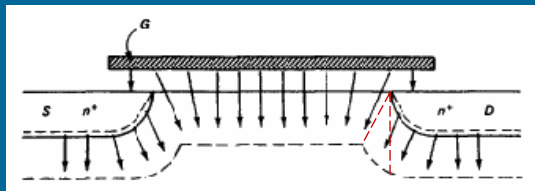
- $E_x < E_y =$ diode
- $E_x > E_y =$ transistor

Well tempered means:

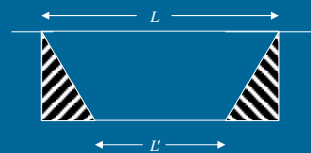
- keeping $E_x > E_y$
- and avoiding the short channel effect

9

SHORT-CHANNEL EFFECT: V_T depends on L



- Charge under gate due to E from G, S and D



- Geometrical construction to estimate V_T drop due to E encroachment

$$\Delta V_T = \frac{\Delta Q_B}{C_{ox}}$$

- Reduce junction depth

10 Pulfrey89

Raised S and D

- Improves I_{ON} by 20-30%

11 Gargini02a

Current and E_y

	1985	2003	200?
VDD, V	5	1	0.8
L, nm	3000	100	15
" E_y ", mV/nm	1.6	10	53

$$I_{Dsat} = \frac{Z}{L} C_{ox} \mu \frac{(V_{GS} - V_T)^2}{2m}$$

m = body factor

$$I_{Dsat} = C_{ox} Z v_{sat} (V_{GS} - V_T)$$

i.e. independent of L
and $f(V_{GS} - V_T)^1$

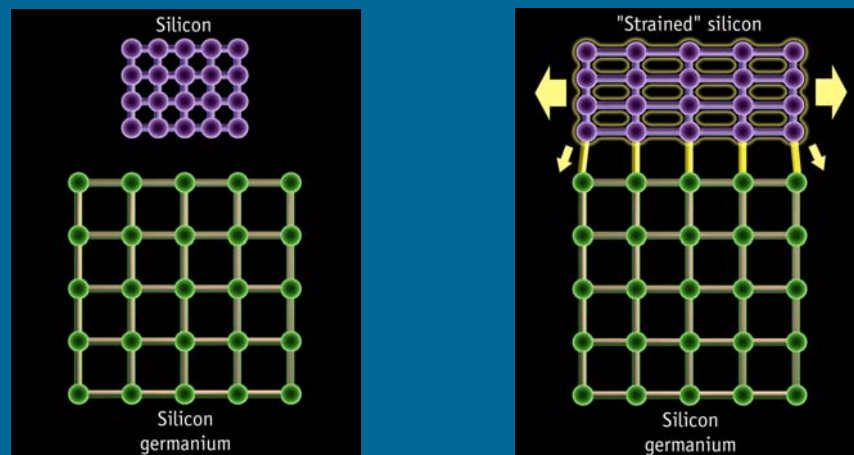
12

MOBILITY

- FETs don't operate at high E_y all the time, or over all of the channel.
- High mobility still very desirable to increase drive current
- Get high μ from strained-silicon channel

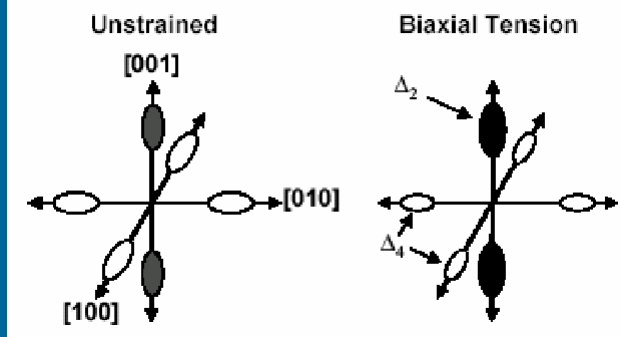
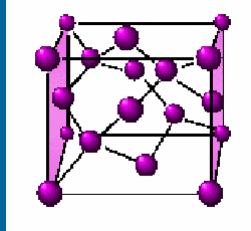
13

Si on SiGe: Tensile strain



14 IBM04

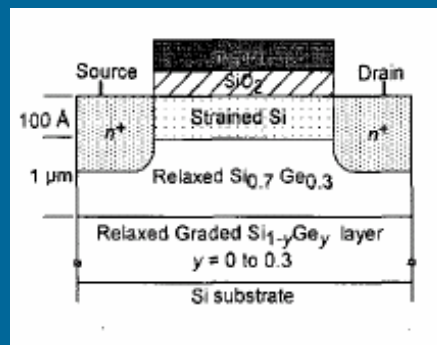
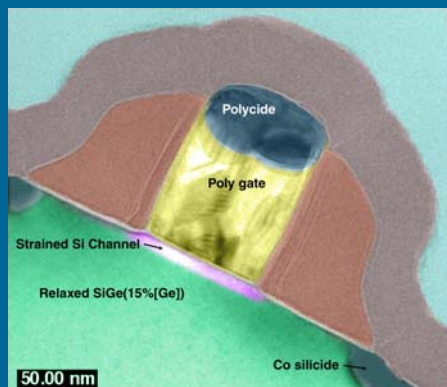
Strained Si: breaking the symmetry



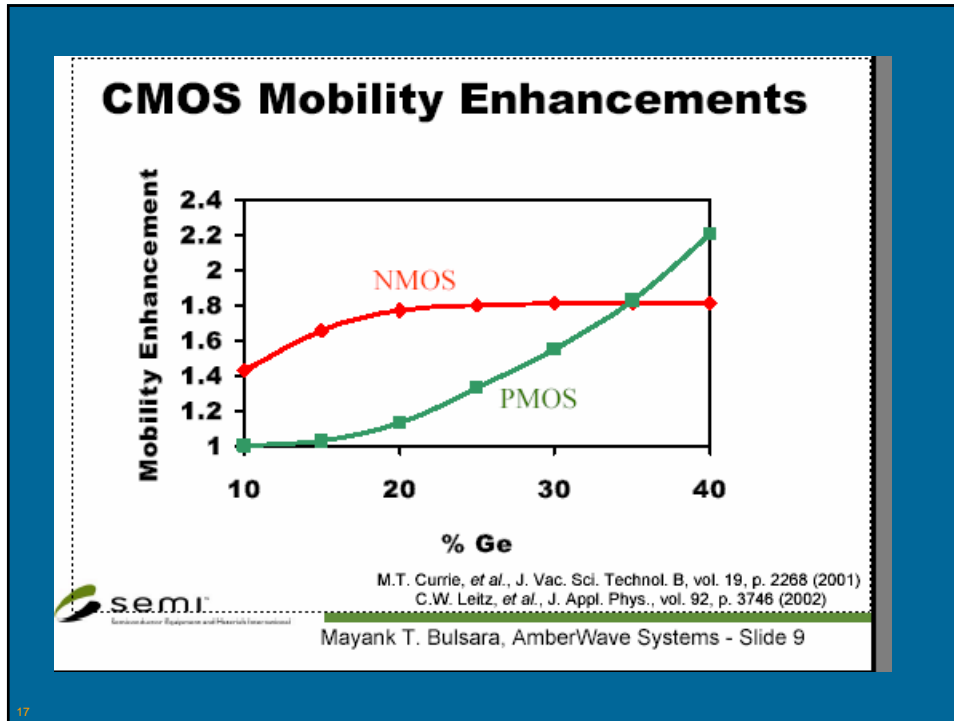
- 6 equivalent directions
- Intervalley scattering
- 2 sub bands lowered in energy
- Reduced intervalley scattering
- Decreased effective mass (horizontal)

15

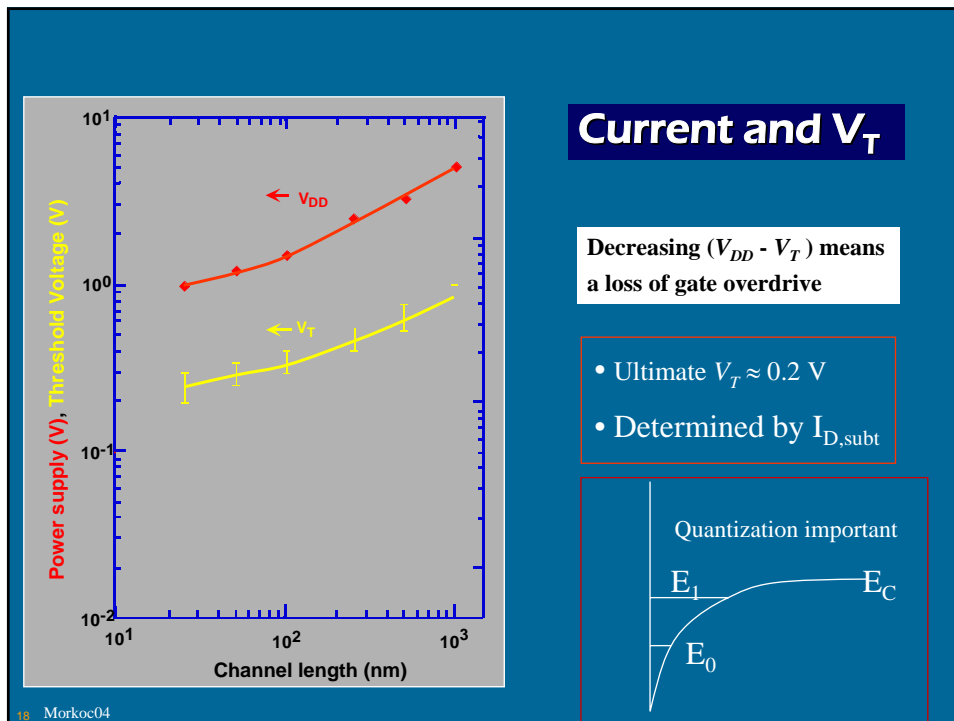
Strained Si: Relaxed sub-layers



16 IBM04



17



Control of $I_{D,subt}$

SOURCE DEPLETION

It's done by capacitive control of the source channel barrier height

$$V_{IB} \equiv \psi_S = \frac{V_{GS}}{1 + \frac{C_B}{C_{ox}}} = \frac{V_{GS}}{m}$$

m: the body factor

$$I_{D,subt}(V_{GS} = 0) = I_{D,thresh} \exp\left[\frac{-V_T}{mV_t}\right]$$

This sets lower limit to V_T , e.g., 0.2V
i.e., $I_{ON}/I_{OFF} \approx 10^4$

19

Sub-Threshold Slope

It's the V_{GS} needed to reduce I_D by 10X

$$S = \left(\frac{d \log_{10} I_D}{dV_{GS}} \right)^{-1} = m 2.303 V_t = m 0.060 V \text{ at } 300 K$$

$$V_t = \frac{kT}{q} \quad \therefore \text{reduce } T$$

Recall: $m = 1 + \frac{C_B}{C_{ox}} = 1 + \frac{\epsilon_s}{\epsilon_{ox}} \cdot \frac{t_{ox}}{W_B}$

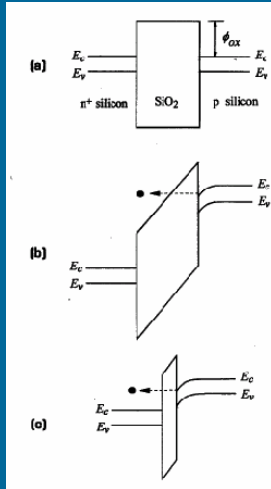
Need: small t_{ox} and small N_A

Gate leakage

V_T compromise

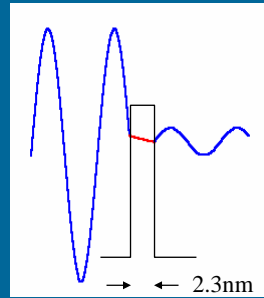
20

Cold-Electron Tunneling



de Broglie wavelength: $\lambda = \frac{h}{mv} = \frac{h}{\sqrt{2m \cdot KE}}$

For an electron in Si at $KE = \phi_{ox}/2$: $\lambda = 6.1 \text{ nm}$



Electron could be either side of the barrier!

(b) FN tunneling
(c) direct tunneling

21 Taur98

Tunneling Facts

$$\text{Tunneling probability } T = \frac{A_{trans}^2}{A_{inc}^2} \approx \exp\left(-\frac{4\pi a}{\lambda}\right) \begin{cases} = 0.0002 \text{ for } 180\text{nm} \\ = 0.0031 \text{ for } 130\text{nm} \\ = 0.0088 \text{ for } 90\text{nm} \end{cases}$$

What is a tolerable gate current?

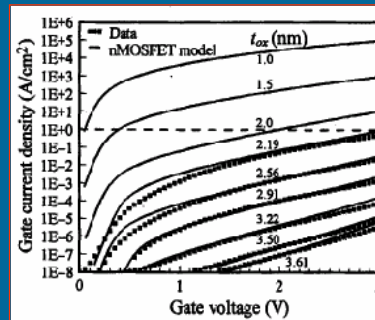
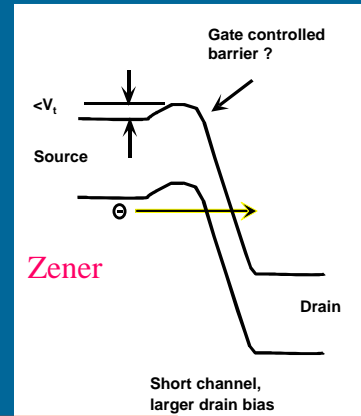
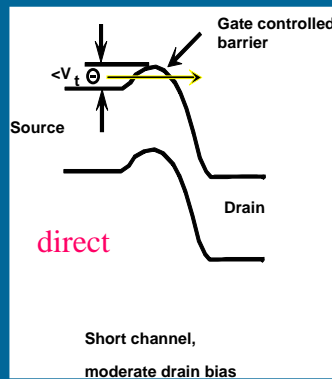


FIGURE 2.44. Measured (dots) and simulated (solid lines) tunneling currents in thin-oxide polysilicon-gate MOS devices. The dashed line indicates a tunneling-current level of 1 A/cm^2 . (After Lo *et al.*, 1997.)

22 Taur98

Ultimate Sub-Threshold Current

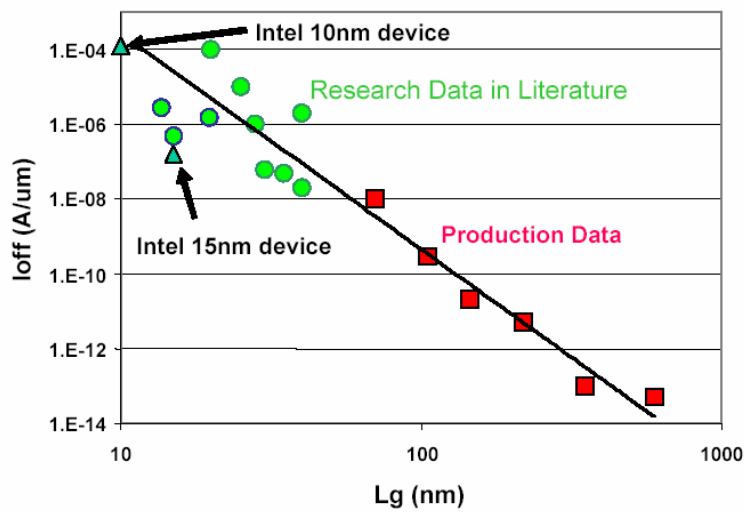
- S → D tunneling



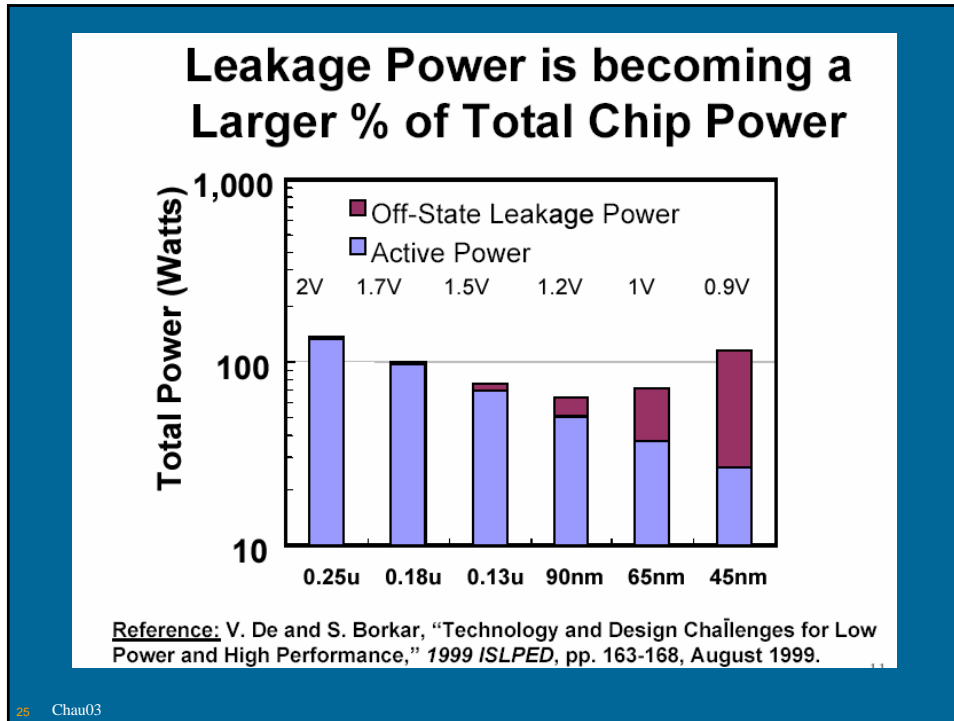
- Expected to occur at $L \approx 10 \text{ nm}$

23 Morkoc04

Transistor Off-state Leakage Trend



24 Chau03



26 Chau03

Power constrained scaling limits

Device type	Application	T (°C)	Power (W/cm ²)	V _{DD} (V)	I _{off} (nA/μm)	V _{Ts} (mV)	t _{oxTeq} (nm)	t _{Si} (nm)	L _{nom} (nm)
Bulk	High performance	85	1000	0.8-1.2	3100-2600	102	0.9-1.0	6-8.5	13-17
		85	100	0.8-1.0	370-340	185	1.1-1.2	8-9	16-18
Bulk	Medium-high performance	85	10	0.6-1.0	50-40	270	1.2-1.4	8-11	16-21
Bulk	Moderate performance	85	1.0	0.6-1.0	6-4.5	360	1.4-1.6	9-12	19-24
Bulk	Low power	65	0.05	0.7-0.9	0.32-0.28	450	1.7-1.8	11-13	24-27
Bulk	Ultralow power	40	<0.001	0.7-1.0	<0.0075	550-710	2.1-2.6	13-19	28-39
Bulk	Moderate-performance SRAM	85	5-1	0.9-1.2	60-10	260-310	1.3-1.6	10-13	20-26
	Low-power SRAM	65	0.1-0.01	0.9-1.2	1.5-0.15	380-470	1.6-2.0	12-16	25-32
	Ultralow-power SRAM	40	0.0001	1.2	0.0018	590	2.4	20	39

It is power dissipation, rather than scaling, that will be the limiting factor *e.g.*, can scale big servers more aggressively than portables and SRAMs

26 Frank02

High-k dielectrics

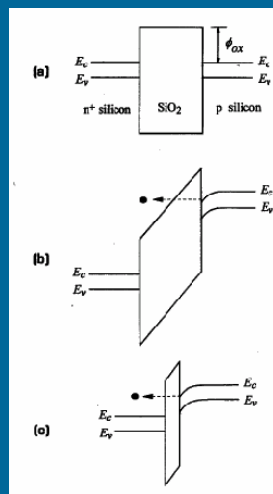
$$C_{ox} = \frac{\epsilon_{ox}}{T_{ox}}$$

- High T_{OX} needed to reduce gate leakage
- High C_{OX} needed for I_D and S
- Resolve conflict by increasing ϵ

Dielectric	Dielectric constant (bulk)
Silicon dioxide (SiO_2)	3.9
Silicon nitride (Si_3N_4)	7
Aluminum oxide (Al_2O_3)	~10
Tantulum pentoxide (Ta_2O_5)	25
Lanthanum oxide (La_2O_3)	~21
Gadolinium oxide (Gd_2O_3)	~12
Yttrium oxide (Y_2O_3)	~15
Hafnium oxide (HfO_2)	~20
Zirconium oxide (ZrO_2)	~23

27 Wong02

High-k dielectrics: tunneling



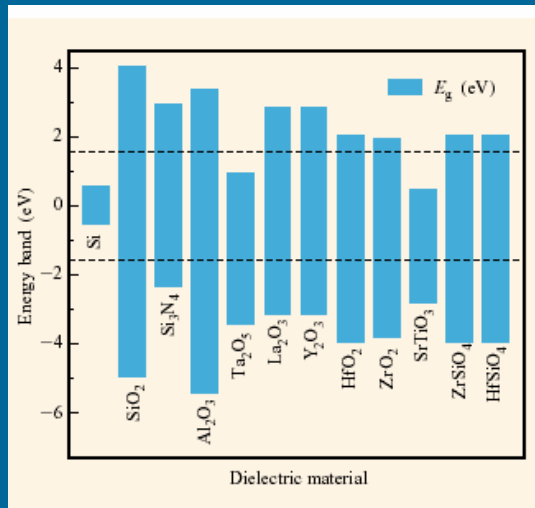
$$\text{Tunneling probability } T = \frac{|A_{trans}|^2}{|A_{inc}|^2} \approx \exp \frac{-4\pi a}{\lambda}$$

$$\text{Tunneling probability } T = \exp \frac{-4\pi}{h} \int_0^a \sqrt{2m [V(x) - E]} dx$$

∴ Need a high ϕ_{ox}

28

High-k dielectrics: contenders



Also:

- must withstand poly activation (950C)
- or use metal gate

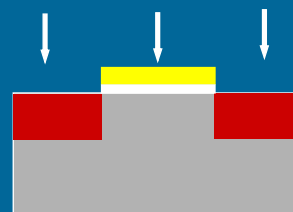
29 Wong02

Metal gate: self-alignment

Poly gates made self alignment possible

Possibilities:

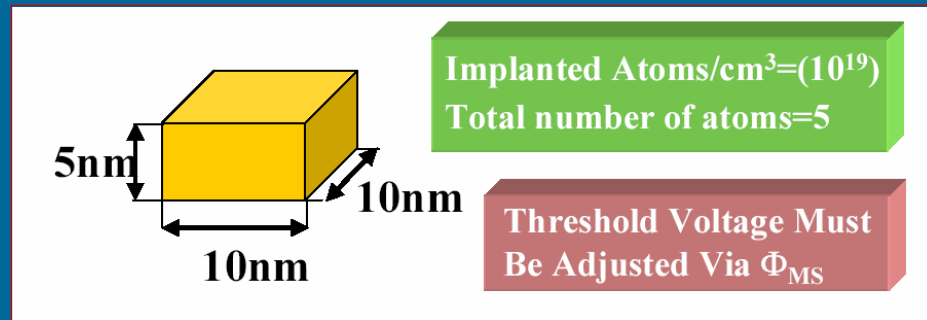
- Perhaps use sacrificial poly gate,
- then deposit metal.
- Co evaporation of metals (Ti and Ni) to obtain different work functions,
- i.e., different V_T 's for NMOS and PMOS or for different blocks on same wafer.



30

Metal gate and N_{SUB}

- If V_T controlled by metal, perhaps can use undoped Si substrate.
- This would remove the problem of dopant fluctuations.



31 Gargini02

Beyond Planar CMOS

Planar CMOS:

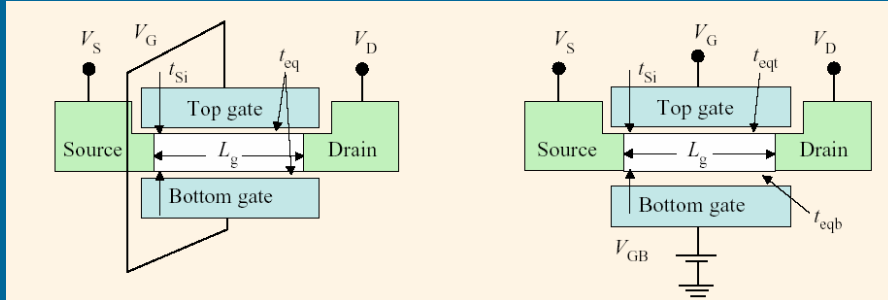
- 10 nm prototypes demonstrated
- raised source and drain
- strained Si
- high kdielectric
- metal gate
- limitation is power dissipation

Further improvements:

- Double gate CMOS
- SOI CMOS

32

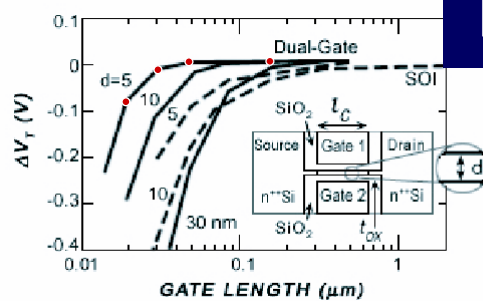
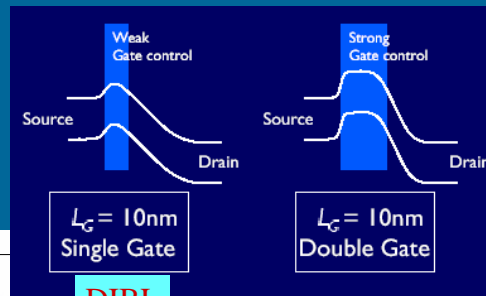
Double-Gate CMOS



- Design flexibility- different V_G 's and T_{ox} 's
- SCE controlled by device geometry, not doping
- Can use undoped channel- reduces statistical fluctuations and Zener BD
- Increased C_{ox} improves I_{ON} and S

33 Wong02

SCE: V_T Roll-off



Note: benefit of shrinking d_{Si}

34 Philips04, MIT02

DG: Improved ON/OFF ratio

Recall : $S = 2.303 m \frac{kT}{q}$

$$m = 1 + \frac{C_B}{C_{ox}}$$

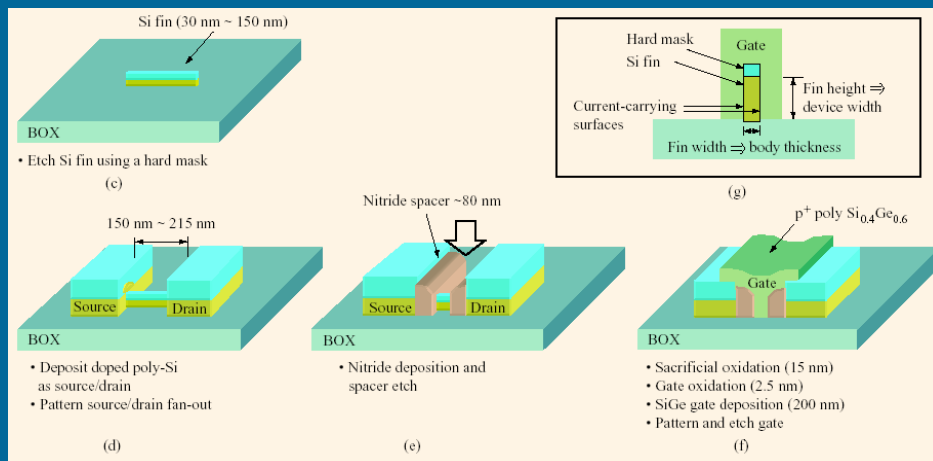
DG doubles without reducing T_{ox}

Tends to zero (small d_{Si} and inversion from top and bottom)

- For same I_{OFF} , set V_T 60mV lower, get more I_{ON}

35

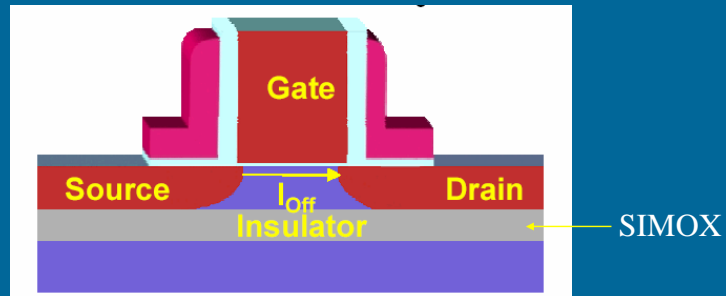
DG example: FINFET



- DG is a deeply scalable FET, but fabrication is difficult

36 Wong02

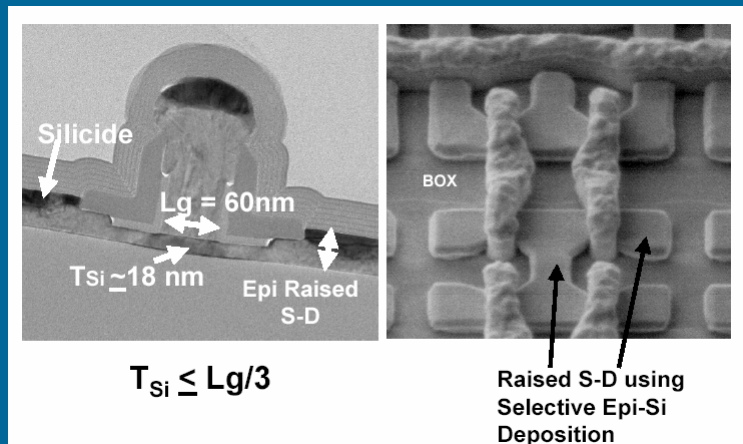
SOI CMOS



- More easily fabricated
- Ultra thin body
- No leakage through substrate
- Very low C_j
- Good device isolation for RF
- Technology of choice for SOC
- Not as deeply scalable as DG

37 Gargini02a

SOI: state-of-the-art



- Small L , x_j , d_{Si}
- Raised S and D
- Fully depleted

38 Chau03

Conclusion



- Planar CMOS: 10nm - THz operation - millions of transistors
- DG CMOS: reduced SCE - best sub threshold slope
- high performance digital
- SOI CMOS: reduced leakage and parasitic C - RF capable
- Do we, or our children, need anything more?

39

References

- Chau03** - <ftp://download.intel.com/research/silicon/Chau%20DRC%20062303%20foils.pdf>
- David04** - ftp://download.intel.com/research/silicon/Ken_David_GSF_030604.pdf
- Frank02** - Frank D.J, IBM J. R&D, v46, 235, 2002
- Gargini02** - <ftp://download.intel.com/research/silicon/PaoloISSUS0102.pdf>
- Gargini02a** - <ftp://download.intel.com/research/silicon/Paolo%20M2S2%200902.pdf>
- IBM03** - <http://www-3.ibm.com/chips/services/foundry/offerings/sige/5hp/>
- IBM04** <http://www.research.ibm.com/resources/press/strainedsilicon/>
- MIT02** <http://ocw.mit.edu/NR/rdonlyres/Electrical-Engineering-and-Computer-Science/6-720JIntegrated-Microelectronic-DevicesFall2002/4E1C74EA-38FB-41CF-B654-3C5AD913B2E9/0/lecture33.pdf>
- Morkoc04** - Morkoc H., WOCSDICE, Slovakia, 2004
- Perlmutter04** - <ftp://download.intel.com/research/silicon/Perlmutter053104.pdf>
- Pulfrey89** - Pulfrey D.L., N.G.Tarr, "Introduction to Microelectronic Devices", Prentice-Hall, 1989
- Raghavan00** - Raghavan G. et al., IEEE Spectrum, v37(10), 47, 2000
- Rucker03** - Rucker R. et al., IEDM, paper 5.3, 2003
- Taur98** - Taur Y., T.H.Ning, "Fundamentals of Modern VLSI Devices", Cambridge University Press, 1998
- Taur99** - Taur Y., IEEE Spectrum, 25, July 1999
- Wong02** - Wong H-S.P., IBM J. R&D, v46, 133, 2002
- Yeo02** - Yeo K-S. et al., "CMOS/BiCMOS VLSI", Prentice-Hall, 2002

40