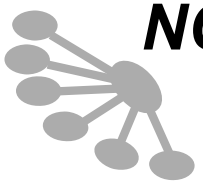
EE-612:
Lecture 17:
Device Scaling

Mark Lundstrom
Electrical and Computer Engineering
Purdue University
West Lafayette, IN USA
Fall 2006



NCN

www.nanohub.org

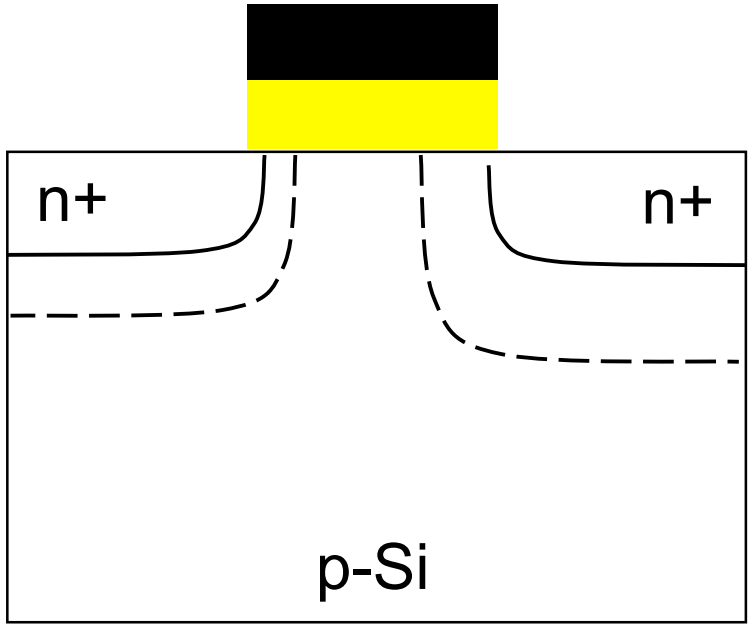
PURDUE UNIVERSITY

outline

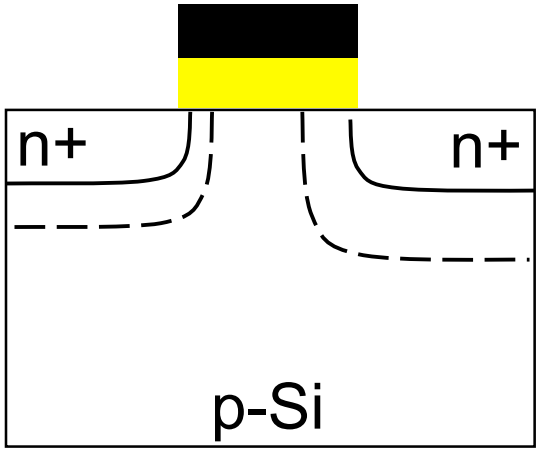
1) Objective of scaling
2) Constant field scaling
3) Non-scaling factors
4) The ITRS
5) Scaling in practice

2D Poisson equation

original device



scaled device

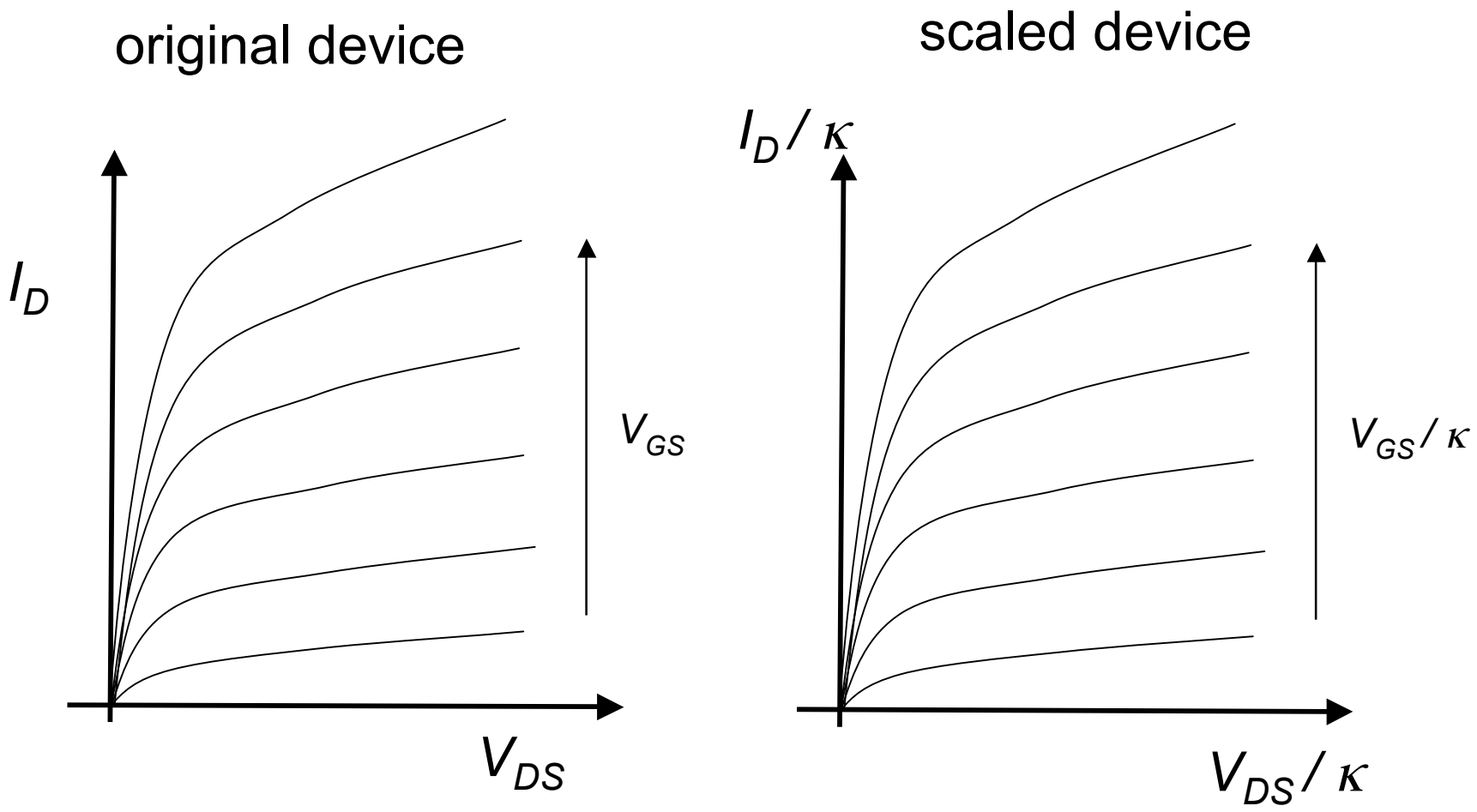


- dimensions reduced by κ
- area reduced by κ^2
- number per chip increased by κ^2

benefits of device scaling

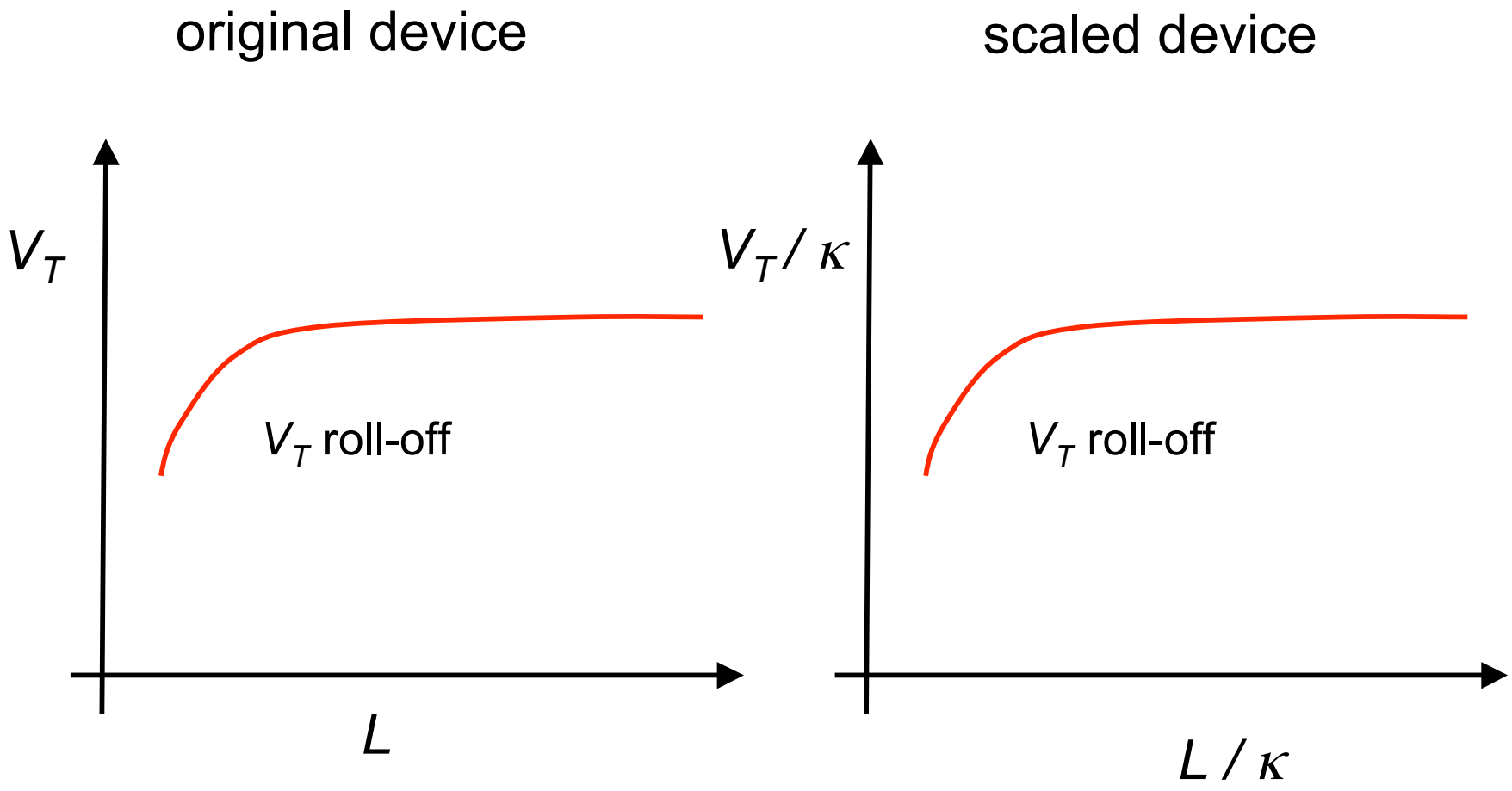
Intel 4004	(1971)	2,250 transistors (L ~ 5-10 microns)	~1MHz
Itanium 2	(2006)	>1,700,000,000 transistors (L ~ 0.065 microns)	~2GHz

scaled I-V



'constant field scaling'

scaled short channel effects



questions

1) How do we scale?
   (e.g. L, t_{ox}, N_A, V_{DD}, etc.)

2) How does performance scale?
   (e.g. *speed, power, etc.)*

3) What does not scale?

4) How does it work in practice?

outline

1) Objective of scaling
2) **Constant field scaling**
3) Non-scaling factors
4) The ITRS
5) Scaling in practice

constant field scaling

Objective:

Maintain a constant electric field as dimensions are scaled down.

$$t_{OX}, L, W, x_j \rightarrow t_{OX}/\kappa, L/\kappa, W/\kappa, x_j/\kappa$$

$$N_A \rightarrow N_A \kappa$$

$$V_{DD} \rightarrow V_{DD}/\kappa$$

constant field scaling (ii)

quantity		***scaled quantity***
E (V/cm)		E
υ (cm/s)	μE	υ
	$\left(\upsilon_{SAT}\right)$	$\left(\upsilon\right)$
$W_D = \sqrt{\frac{2\varepsilon_{Si}}{qN_A}\left(V_{bi} + V_{DD}\right)}$ (cm)		W_D / κ
$C = \frac{\varepsilon A}{t}$ (F)		C / κ

constant field scaling (iii)

quantity	***scaled quantity***
$Q_i = C_{OX}\left(\mathrm{V_{GS}} - V_T\right)(\mathrm{C/cm^2})$	Q_i
$I_D = W\,Q_i\,\upsilon \quad (\mathrm{A})$	I_D / κ
$R_{CH} = \dfrac{V_{DS}}{I_D} = \dfrac{L}{W\,\mu_{eff} Q_i} \quad (\Omega)$	R_{CH}
$V_T = V_{FB} + 2\psi_B + \sqrt{2\mathrm{q}\varepsilon_{\mathrm{Si}} N_A\left(2\psi_B\right)}\Big/ C_{OX} \quad (\mathrm{V})$	$> V_T/\sqrt{\kappa}$
	need V_T/κ

impact on circuits / systems

quantity	***scaled quantity***
$\tau = CV_{DD}/I_D \quad (\text{sec})$	τ/κ
$P = V_{DD}I_D \quad (\text{W})$	P/κ^2
$D = \frac{\text{no.}}{A_C} \quad (\text{cm}^{-2})$	$D \times \kappa^2$
$P/A \quad (\text{W/cm}^2)$	P/A
$P\tau = CV_{DD}^2$	$P\tau/\kappa^3$

constant voltage scaling

Objective:

Maintain a constant power supply voltage as dimensions are scaled down.

$$V_{DD} \rightarrow V_{DD}$$

$$t_{OX}, L, W, W_D, x_j \rightarrow t_{OX}/\kappa, L/\kappa, W/\kappa, W_D/\kappa, x_j/\kappa$$

$$N_A \rightarrow N_A \kappa^2$$

electric fields increase!!
power density increases!!

$$W_D = \sqrt{\frac{2\varepsilon_{Si}}{qN_A}\left(V_{bi} + V_{DD}\right)} \text{ (cm)}$$

generalized scaling

$$L, t_{ox}, \text{etc.} \rightarrow L/\kappa, t_{ox}/\kappa, \text{etc.}$$

$$E \rightarrow \alpha E$$

$$V_{DD} \rightarrow \alpha V_{DD}/\kappa > V_{DD}/\kappa$$

constant field: $\alpha = 1$
constant voltage: $\alpha = \kappa$

See Taur and Ning, pp. 168-169

impact on circuits / systems

quantity	***scaled quantity***
$\tau = CV_{DD}/I_D \quad (\text{sec})$	τ/κ
$P = V_{DD}I_D \quad (\text{W})$	P/κ^2
$D = \frac{\text{no.}}{A_C} \quad (\text{cm}^{-2})$	$D \times \kappa^2$
$P/A \quad (\text{W/cm}^2)$	P/A
$P\tau = CV_{DD}^2$	$P\tau/\kappa^3$

generalized scaling

$$L, t_{ox}, \text{etc.} \rightarrow L/\kappa, t_{ox}/\kappa, \text{etc.}$$

$$E \rightarrow \alpha E$$

$$V_{DD} \rightarrow \alpha V_{DD}/\kappa > V_{DD}/\kappa$$

(constant field: $\alpha = 1$
constant voltage: $\alpha = \kappa$)

See Taur and Ning, pp. 168-169

outline

1) Objective of scaling
2) Constant field scaling
3) **Non-scaling factors**
4) The ITRS
5) Scaling in practice

bandgap, ψ_B

$$V_T = V_{FB} + 2\psi_B + \sqrt{2q\varepsilon_{Si}N_A(2\psi_B)}\Big/C_{OX} \quad \text{(V)}$$

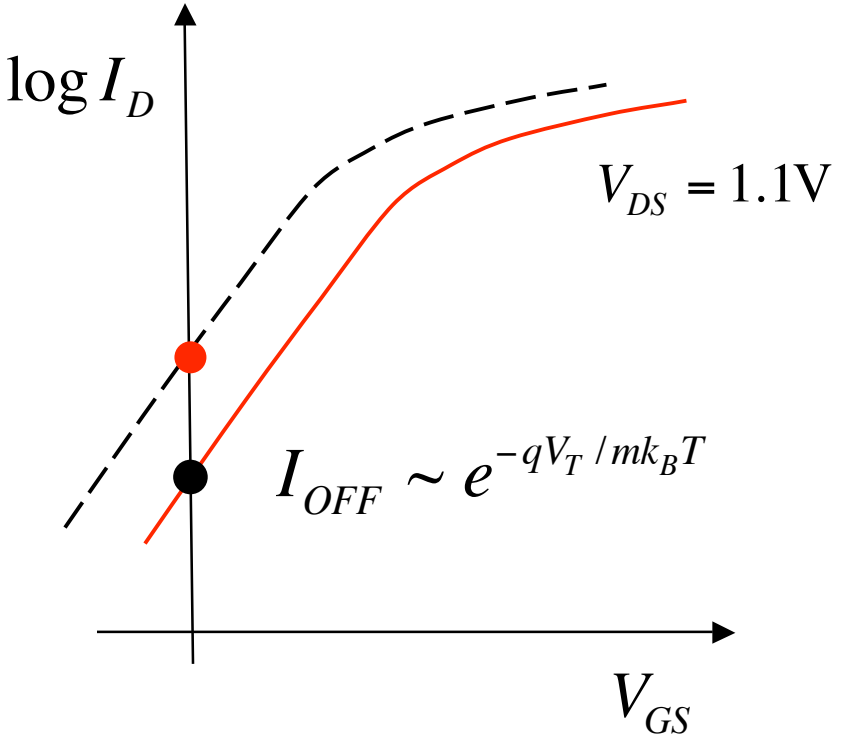
$$V_{FB} = -\frac{E_G}{2q} - \psi_B$$

$$\psi_B = \frac{k_B T}{q}\ln\left(\frac{N_A}{n_i^2}\right)$$

$$V_T \rightarrow V_T/\kappa$$

will take some work!

subthreshold behavior



$$m = 1 + C_D/C_{OX}$$

_m_ does not scale

_S_ does not scale!!

I_{OFF} increases

I_{OFF} spec sets minimum V_T

V_T spec sets minimum V_{DD}

$$I_{ON} \sim (V_{DD} - V_T)$$

V_{DD} scaling has stopped at about 1.1V

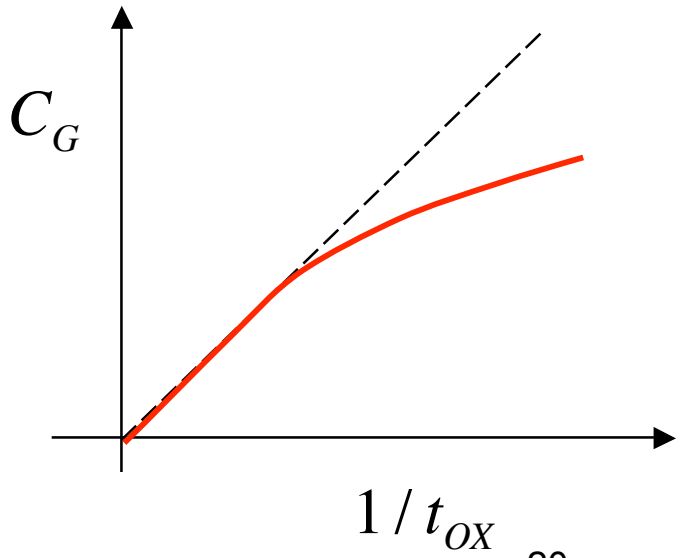
inversion layer thickness

$$t_{inv} = \frac{2k_BT/q}{E_S} = \frac{2\varepsilon_{Si}k_BT/q}{Q_i}$$

Q_i does not scale for constant field scaling, so…..

t_{inv} does not scale

$$C_G = \left(\frac{1}{C_{OX}} + \frac{1}{C_S}\right)^{-1} = \left(\frac{t_{OX}}{\varepsilon_{OX}} + \frac{t_{inv}}{\varepsilon_{Si}}\right)^{-1}$$



effective mobility

$$\mu_{eff} \approx \mu_{SR} \cdot E_{eff}^{-\eta}$$

$$E_{eff} = \frac{1}{\varepsilon_{Si}}\left(|Q_{DM}| + \frac{|Q_i|}{2}\right)$$

increases with scaling

μ_{eff} decreases with scaling

outline

1) Objective of scaling
2) Constant field scaling
3) Non-scaling factors
4) **The ITRS**
5) Scaling in practice

ITRS

"The International Technology Roadmap for Semiconductors, known throughout the world as the ITRS, is the fifteen-year assessment of the semiconductor industry's future technology requirements. These future needs drive present-day strategies for world-wide research and development among manufacturers' research facilities, universities, and national labs."

http://www.itrs.net/

http://www.itrs.net/

ITRS 2005 Edition

- Executive Summary
- System Drivers
- Design
- Test & Test Equipment
- **Process Integration, Devices & Structures**
- RF and A/MS Technologies for Wireless Communications
- Emerging Research Devices (includes Emerging Research Materia
- Front End Processes
- Lithography
- Interconnect
- Factory Integration
- Assembly & Packaging
- Environment, Safety & Health
- Yield Enhancement
- Metrology
- Modeling & Simulation

Acronyms

ITRS: PIDS (near term)

Table 40a High-Performance Logic Technology Requirements—Near-term

Grey cells delineate one of two time periods: either before initial production ramp has started for ultra-thin body fully depleted (UTB FD) SOI or double-gate (DG) MOSFETs, or beyond when planar bulk or UTB FD MOSFETs have reached the limits of practical scaling (see the text and the table notes for further discussion)

Year of Production	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*
DRAM ½ Pitch (nm) (contacted)	*80*	*70*	*65*	*57*	*50*	*45*	*40*	*36*	*32*
MPU/ASIC Metal 1 (M1) ½ Pitch (nm)(contacted)	*90*	*78*	*68*	*59*	*52*	*45*	*40*	*36*	*32*
MPU Physical Gate Length (nm)	*32*	*28*	*25*	*22*	*20*	*18*	*16*	*14*	*13*
L_g: Physical L_{gate} for High Performance logic (nm) [1]	32	28	25	22	20	18	16	14	13
EOT: Equivalent Oxide Thickness [2]									
Extended planar bulk (Å)	12	11	11	9	7.5	6.5	5	5	
UTB FD (Å)				9	8	7	6	5	5
DG (Å)							8	7	6
Gate Poly Depletion and Inversion-Layer Thickness [3]									
Extended Planar Bulk (Å)	7.3	7.4	7.4	2.9	2.8	2.7	2.5	2.5	
UTB FD (Å)				4	4	4	4	4	4
DG (Å)							4	4	4
EOT_{elec}: Electrical Equivalent Oxide Thickness in inversion [4]									
Extended Planar Bulk (Å)	19.3	18.4	18.4	11.9	10.3	9.2	7.5	7.5	
UTB FD (Å)				13	12	11	10	9	9
DG (Å)							12	11	10
$J_{g,limit}$: Maximum gate leakage current density [5]									
Extended Planar Bulk (A/cm^2)	1.88E+02	5.36E+02	8.00E+02	9.09E+02	1.10E+03	1.56E+03	2.00E+03	2.43E+03	
FDSOI (A/cm^2)				7.73E+02	9.50E+02	1.22E+03	1.38E+03	2.07E+03	2.23E+03
DG (A/cm^2)							6.25E+02	7.86E+02	8.46E+02
V_{dd}: Power Supply Voltage (V) [6]	1.1	1.1	1.1	1	1	1	1	0.9	0.9

ITRS: PIDS (long term)

Table 40b High-Performance Logic Technology Requirements—Long-term

Grey cells delineate one of two time periods: either before initial production ramp has started for ultra-thin body fully depleted (UTB FD) SOI or double-gate (DG) MOSFETs, or beyond when planar bulk or UTB FD MOSFETs have reached the limits of practical scaling (see the text and the table notes for further discussion).

Year of Production	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
DRAM ½ Pitch (nm) (contacted)	*28*	*25*	*22*	*20*	*18*	*16*	*14*
MPU/ASIC Metal 1 (M1) ½ Pitch (nm)(contacted)	*28*	*25*	*22*	*20*	*18*	*16*	*14*
MPU Physical Gate Length (nm)	*11*	*10*	*9*	*8*	*7*	*6*	*6*
L_g: *Physical* L_{gate} *for High Performance logic (nm) [1]*	11	10	9	8	7	6	5
EOT: Equivalent Oxide Thickness [2]							
Extended planar bulk (Å)							
UTB FD (Å)	5	5					
DG (Å)	6	6	5	5	5	5	5
Gate Poly Depletion & Inversion-Layer Thickness [3]							
Extended planar bulk (Å)							
UTB FD (Å)	4	4					
DG (Å)	4	4	4	4	4	4	4
EOT_{elec}: *Electrical Equivalent Oxide Thickness in inversion [4]*							
Extended Planar Bulk (Å)							
UTB FD (Å)	9	9					
DG (Å)	10	10	9	9	9	9	9
$J_{g,limit}$: *Maximum gate leakage current density [5]*							
Extended Planar Bulk (A/cm^2)							
FDSOI (A/cm^2)	3.27E+03	3.70E+03					
DG (A/cm^2)	1.00E+03	1.10E+03	1.22E+03	1.38E+03	1.57E+03	1.83E+03	2.20E+03
V_{dd}: *Power Supply Voltage (V) [6]*	0.9	0.8	0.8	0.7	0.7	0.7	0.7

outline

1) Objective of scaling
2) Constant field scaling
3) Non-scaling factors
4) The ITRS
5) **Scaling in practice**

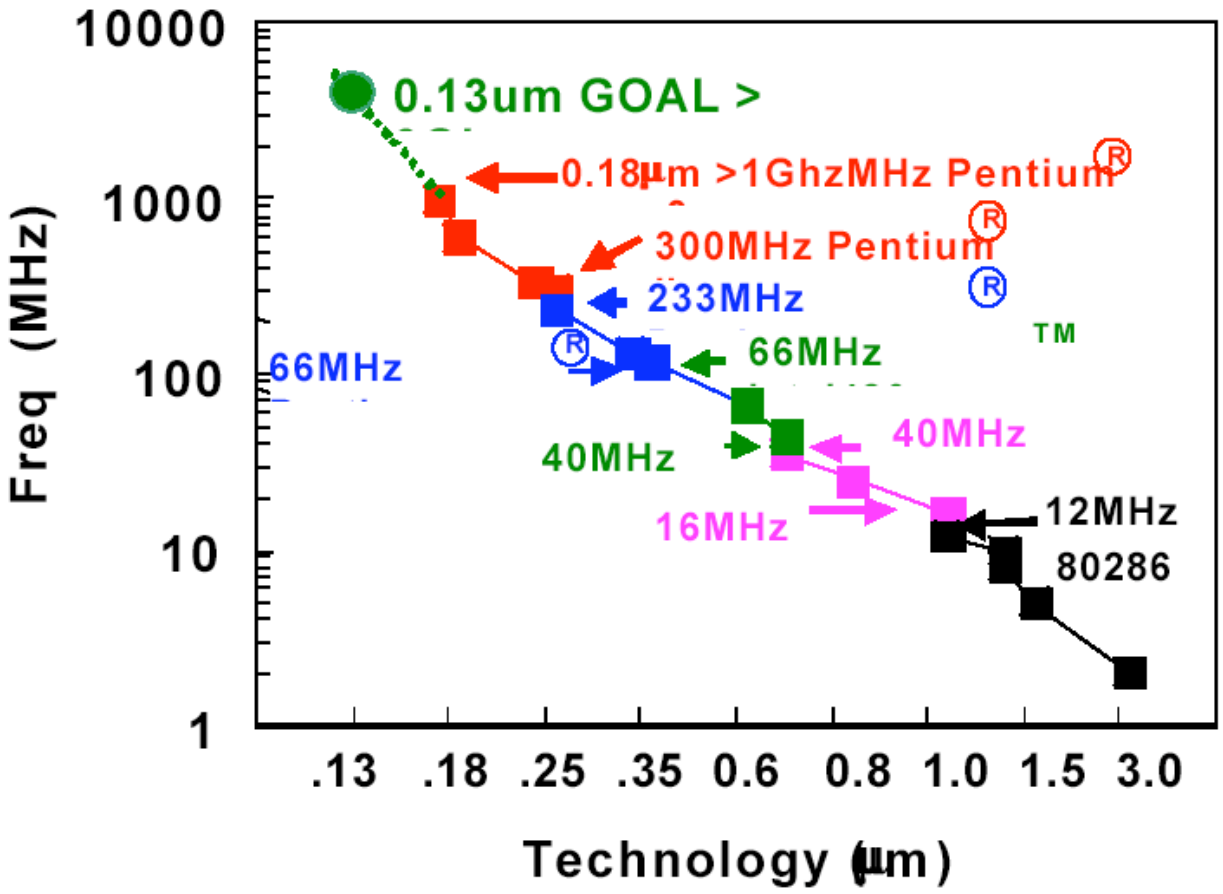
scaling in practice

130nm Logic Technology Featuring 60nm Transistors, Low-K Dielectrics, and Cu Interconnects
Scott Thompson, et al.,

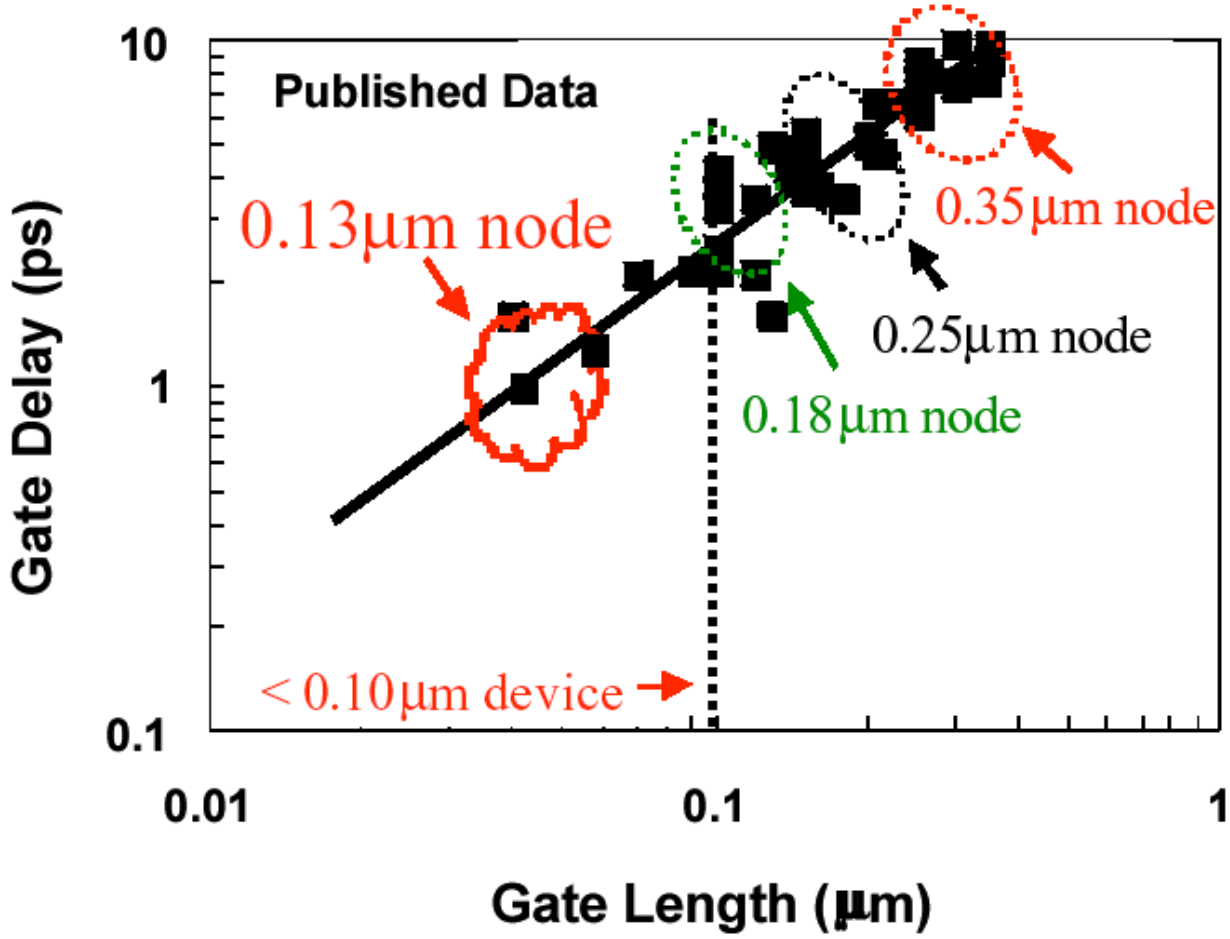
Transistor Elements for 30nm Physical Gate Lengths and Beyond
Brian Doyle, et al.

Intel Technology Journal, Vol. 6 Issue 2, 2002

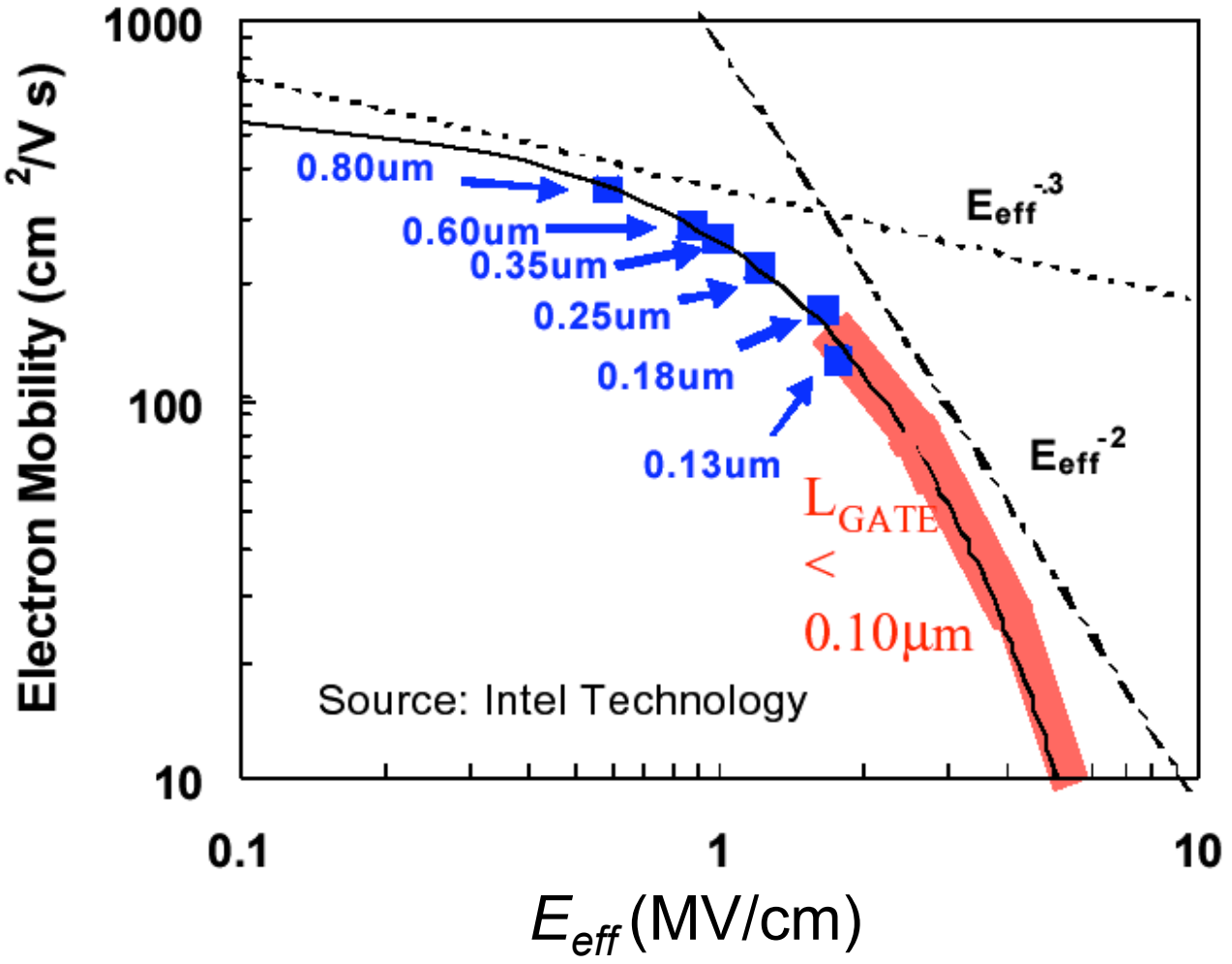
speed vs. scaling



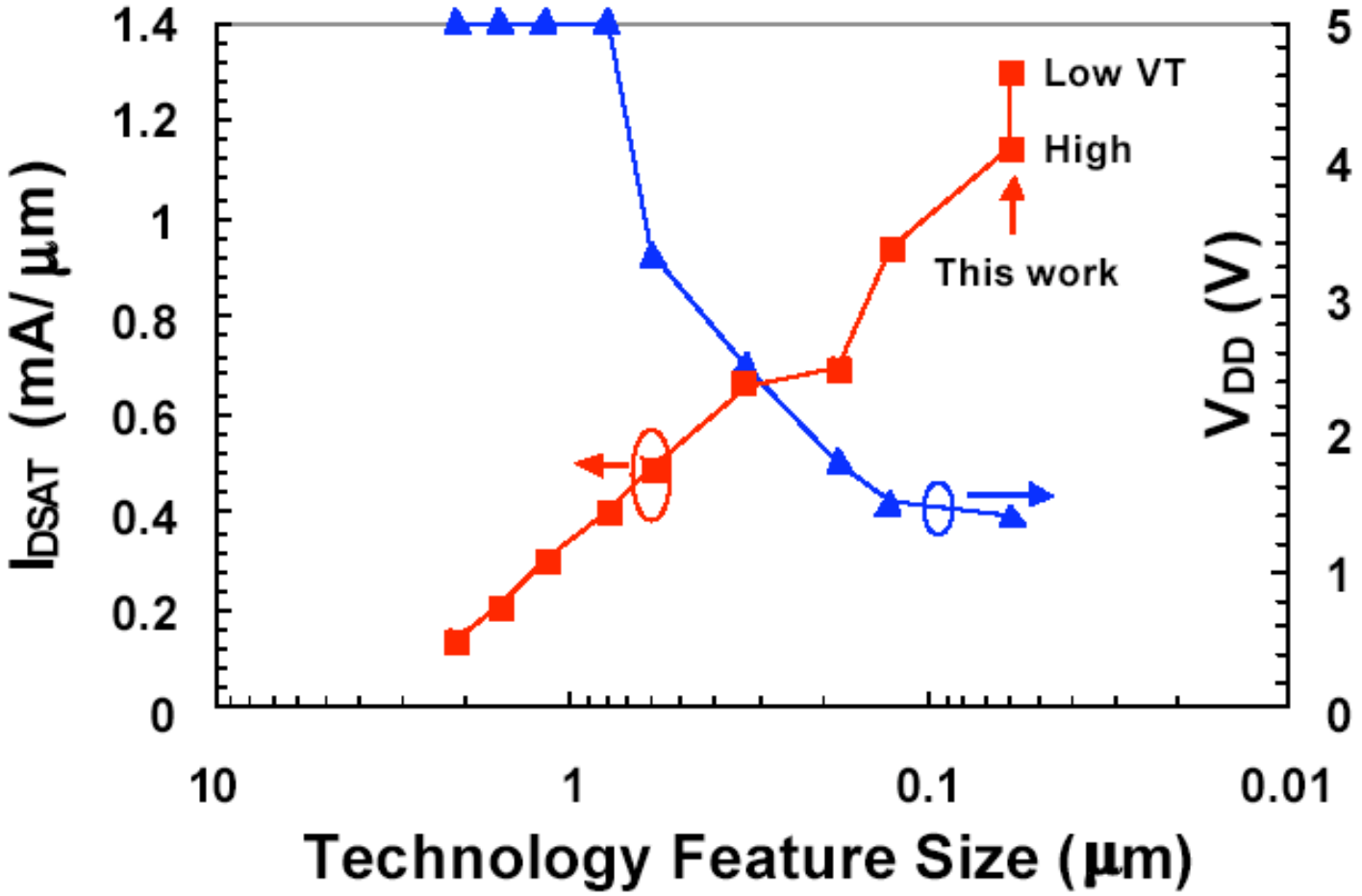
device delay vs. scaling



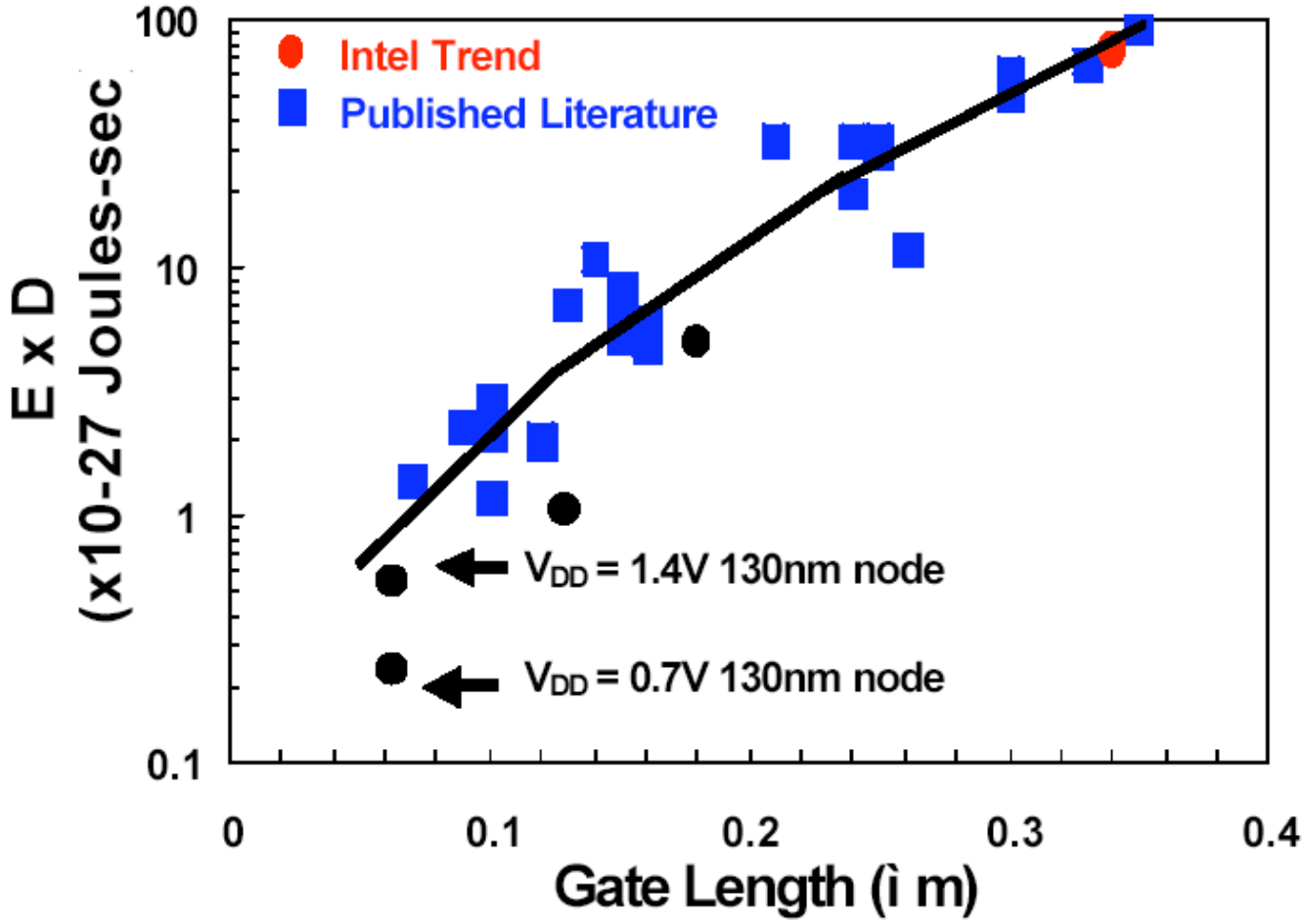
effective mobility vs. scaling



I_{ON} and V_{DD} vs. scaling



energy-delay product vs. scaling



outline

1) Objective of scaling
2) Constant field scaling
3) Non-scaling factors
4) The ITRS
5) Scaling in practice