

CMOS Technology Scaling

- See page 122-129 of Rabaey text
- Full Scaling – V_{DD} scaled by S , Dimensions by S
- General Scaling - V_{DD} scaled by U , Dimensions by S
- Fixed-Voltage Scaling - only Dimensions by S
- Overall Capacitance scales by $1/S$, all scaling models
- Delay by $1/S$ (short channel devices, all scaling models)
- With deep sub-micron, fixed voltage scaling is now the rule

BR 6/00

1

MOSFET

- Four terminals – G, S, D, Bulk (substrate)
- Three regions of operation
 - Cutoff ($V_{GS} < V_t$)
 - Linear (Resistive) - drain current linear with increasing V_{DS} , $V_{GS} > V_t$
 - Saturation (drain current constant with increasing V_{DS}), $V_{GS} > V_t$
- Behavior of long channel devices ($> 1.0 \mu\text{m}$) significantly different from short channel devices ($< 0.5 \mu\text{m}$).
- Transistors in digital circuits spent most of their time in either cutoff or saturation, so we are principally interested in those regions.

BR 6/00

2

Regions of Operation

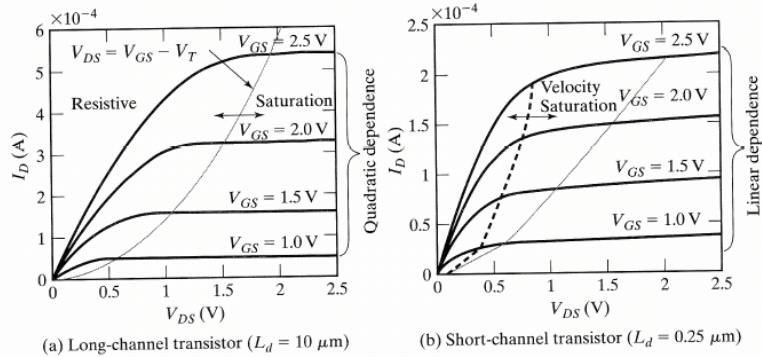


Figure 3-19 I - V characteristics of long- and a short-channel NMOS transistors in a $0.25 \mu\text{m}$ CMOS technology. The (W/L) ratio of both transistors is identical and equals 1.5. Observe the difference in the y -axis scale.

BR 6/00

3

Long Channel Scaling, Delay

Drain Current, MOS transistor, in Saturation, Long Channel

$$I_{\text{DSAT}} = K'_n / 2 * W/L * (V_{\text{GS}} - V_{\text{T}})^2 (1 + \lambda * V_{\text{DS}}) \quad \text{pg. 93}$$

Recall that $K'_n = \mu_n C_{\text{ox}}$, so will scale by S because C_{ox} increases with decreasing thickness. If K'_n scales by S , then so will I_{D} .

C_L is gate capacitance = $C_{\text{ox}} * W * L$

Scales by $1/S$ because both W , L scale by $1/S$ but C_{ox} scales by S (C_{ox} increases with decreasing thickness).

Intrinsic Delay = $C_L * V / I_{\text{D}}$; so will scale by $1/S^2$!!

BR 6/00

4

Short Channel, Velocity Saturation

The velocity of carriers in the channel can be expressed as:

$$v_n = -\mu_n E(x) = \mu_n dV/dx \quad \text{pg 92.}$$

μ_n is the electron mobility. Simply put, the stronger the electric field across the channel, the higher the velocity (and faster the device).

There is a limit though. When $E(x)$ reaches a critical value E_{sat} , the velocity of the carriers saturate (Figure 3-17, pg 94, textbook).

For p-type silicon (NMOS transistor), $E_{sat} = 1.5V/\mu m$

For channel length = $0.25\mu m$, only need $V_{DS} = 2V$

BR 6/00

5

Short Channel I_{DSAT} Equation

An approximation for I_{DSAT}

$$I_{DSAT} = v_{SAT} C_{ox} W (V_{GS} - V_T - (V_{DSAT}/2)) \quad \text{pg 97.}$$

V_{DSAT} is drain-source voltage when velocity saturation occurs.

Saturation current now has linear dependence on V_{GS} (instead of squared). Reducing the operating voltage does not have as much effect on short-channel devices as in long-channel devices.

I_{DSAT} is independent on L . I_{DSAT} scaling is constant for constant voltage scaling since C_{ox} scales by $1/S$ and W by S .

BR 6/00

6

MOSFET Capacitance

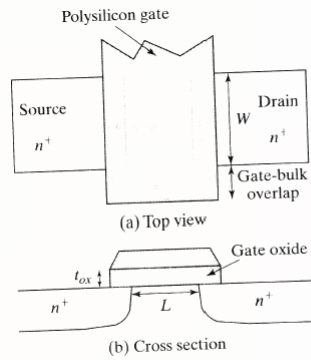


Figure 3-29 MOSFET overlap capacitance.

C_{gc} – gate to channel

Has three components:

C_{gcs} – gate to source (overlap)

C_{gcd} – gate to drain (overlap)

C_{gcb} – gate to channel

BR 6/00

7

Capacitance varies with operation

Table 3-4 Average distribution of channel capacitance of MOS transistor for different operation regions.

Operation Region	C_{GCB}	C_{GCS}	C_{GCD}	C_{GC}	C_G
Cutoff	$C_{ox}WL$	0	0	$C_{ox}WL$	$C_{ox}WL + 2C_oW$
Resistive	0	$C_{ox}WL/2$	$C_{ox}WL/2$	$C_{ox}WL$	$C_{ox}WL + 2C_oW$
Saturation	0	$(2/3)C_{ox}WL$	0	$(2/3)C_{ox}WL$	$(2/3)C_{ox}WL + 2C_oW$

Due to pinch-off of channel during saturation

BR 6/00

8

Short Channel Scaling, Delay

$$\text{Intrinsic Delay} = C_L * V / I_D$$

C_L scales by $1/S$, I_D will be constant for constant voltage scaling, so delay only scales by $1/S$.

Another problem is that electron mobility degrades with short channel devices as well. This will also decrease the delay scaling for short channel devices.

BR 6/00

9

Power Scaling

$$P_{av} = C_L * V^2 / T_p$$

where T_p is intrinsic delay.

For long channel devices, this scales by S (constant voltage)

For short channel devices, scaling is 1 (constant voltage).

What about power density??? (power per unit area).
Even with P_{av} being '1' for short channel devices, if we had N devices before in a given area, we can now pack N^2 devices in the same area since W, L are both scaled by $1/S$.

So power density scales by S^2 !!!!!

BR 6/00

10

Wire Resistance Scaling

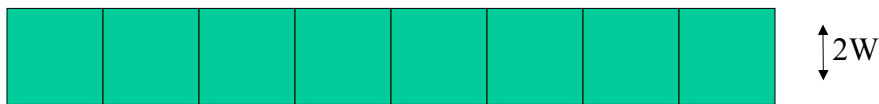
Wire Resistance = ohms/square * L / W where

ohms/square is a constant that depends on resistivity of material of the wire = R_{sq}



$$R_{\text{wire}} = R_{sq} * L / W$$

What if we double the wire width and keep the same L?



$$R_{\text{wire_new}} = R_{sq} * L / 2W = R_{\text{wire old}} / 2$$

BR 6/00

11

Wire Resistance Scaling (cont)

Wiring width always scales by 1/S

Wiring length scales differently depending upon whether it is global wiring or local wiring.

Global wiring spans the chip, and die sizes are remaining constant to increasing. L for global wiring remains constant.

Local wiring on spans a region. L for global wiring scales by 1/S

R_{wire} is constant for local wiring (both L, W decrease).

R_{wire} scales by S by global wiring since W decreases but L remains the same.

BR 6/00

12

Sheet Resistance (Leda 0.25U)

	Sheet Res (ohms/sq)
N+	4.9
P+	3.5
Poly	4.2 (silicided to reduce resistance)
Metal Layers	0.07

Aluminum Resistivity: $2.65 \text{ e-}8$ Ohm-meters

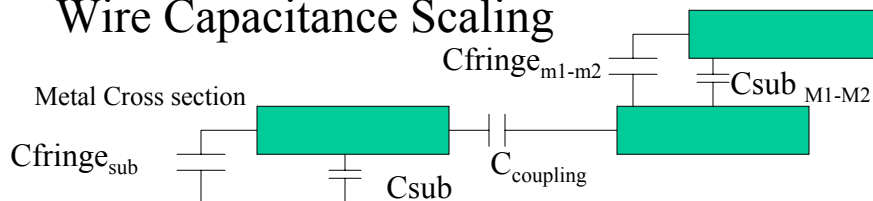
Copper Resistivity: $1.67 \text{ e-}8$ Ohm-meters

Resistivity of Aluminum about 60% higher than copper. Copper interconnect preferred – more expensive fabrication

BR 6/00

13

Wire Capacitance Scaling



$$C_{\text{wire}} = C_{\text{fringe}} + C_{\text{sub}} + C_{\text{coupling}}$$

$$C_{\text{sub}} = C_{\text{ox}} * W * L$$

Recall that C_{ox} scales by S . So C_{sub} scales by $1/S$.

C_{fringe} depends on thickness of sidewall, and L of wire, C_{ox} of insulator. Thickness will remain constant.

$C_{\text{fringe}}(\text{sub})$ will be constant (L scales by $1/S$, C_{ox} by S).

C_{coupling} is *controlled via spacing rules*. In sub-micron technologies, minimum spacing often controlled by capacitance considerations.

BR 6/00

14

Interconnect Capacitance (Substrate)

	Poly	M1	M2	M3	M4	M5
Sub	113	37	18	13	9	8
poly		53	16	10	7	6
m1			35	15	9	7
m2				39	16	10
m3					44	16
m4						39

Numbers for Leda 0.25u. Units = af/ μm^2

Note that units are based on Area.

BR 6/00

15

Interconnect Capacitance (Fringe)

	M1	M2	M3	M4	M5
Sub	21	60	56	40	25
poly	70	39	30	25	22
m1		62	36	28	24
m2			61	38	30
m3				55	39
m4					62

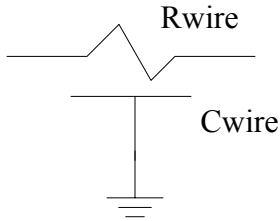
Numbers for Leda 0.25u. Units = af/ μm

Note that units are based only on Length of wire.

BR 6/00

16

Wire Delay



If C_{sub} dominates, the C_{wire} scales by $1/S$
 R_{wire} is constant for local wires, so local wire $R_{wire} * C_{wire}$ (delay) scales by $1/S$ which is good news (gate delay also scales by $1/S$).

R_{wire} scales by S for global wires, so global wire delay $R_{wire} * C_{wire}$ (delay) is constant!!! The gate delays scale down, so global wire delay scales UPWARD with respect to gate delays. BAD!!

BR 6/00

17

Clock Speed Scaling

Most systems have less than 16 FO4 delays between registers.

System clock speed determined by:

Clock2Q of Register + Register2Register Delay + Setup + clock Skew budget

Clock2Q, Setup, Register2Register delay scales down with technology.

However, clock is a global signal. Clock skews remain constant, and grow relative to gate delays. This means that more and more of the clock period is taken up by clock skew budget. Have to solve this by clever design techniques, local clocks, matching of data delays to clock skew delays.

BR 6/00

18

Clock Evolution in Alpha Microprocessor

Alpha 21064 (0.75u to 0.25u), clock from 150 Mhz to 275Mhz.

One large clock driver, 3.5nF load, about 160ps skew across chip (clock skew 4.4% of 200 Mhz clock cycle)

Alpha 21164 (0.5u) – clock from 300 Mhz to 366 Mhz. Multiple clock buffering via tree, but still only 1 clock. Clock buffering reduced skew to 80 ps. Clock skew now 3% of clock design due to new clock design.

Alpha 21264 (0.35u) – clock up to 600 Mhz. Used local clocking to save power, max skew was 72ps. Clock Skew is now 4.3% of clock period.

Economic Scaling

- Advanced Fabs keep getting more and more expensive.
- New Fabrication line cost on order of low Billions for <0.15u
 - Partnerships between companies
- Masks cost go up as well
- NRE becomes extremely high – will either have to produce LOTS of one design or re-used actual chips
 - Reconfigurable hardware will become increasingly important due to economics.

High Performance – Near Term Future Tech Requirements

Table 35a High-performance Logic Technology Requirements—Near-term

YEAR OF PRODUCTION	2001	2002	2003	2004	2005	2006	2007
DRAM 1/2 PITCH (nm)	130	115	100	90	80	70	65
MPU / ASIC 1/2 PITCH (nm)	150	130	107	90	80	70	65
MPU PRINTED GATE LENGTH (nm)	90	75	65	53	45	40	35
MPU PHYSICAL GATE LENGTH (nm)	65	53	45	37	32	28	25
Physical gate length high-performance (HP) (nm) [1]	65	53	45	37	32	28	25
Equivalent physical oxide thickness for high-performance T_{ox} (EOT) (nm) [2]	1.3–1.6	1.2–1.5	1.1–1.6	0.9–1.4	0.8–1.3	0.7–1.2	0.6–1.1
Gate depletion and quantum effects electrical thickness adjustment factor (nm) [3]	0.8	0.8	0.8	0.8	0.8	0.8	0.5
T_{ox} electrical equivalent (nm) [4]	2.3	2.1	2.0	2.0	1.9	1.9	1.4
Nominal power supply voltage (V_{dd}) (V) [5]	1.2	1.1	1.0	1.0	0.9	0.9	0.7
Nominal high-performance NMOS sub-threshold leakage current, I_{dsub} (at 25°C) ($\mu A/\mu m$) [6]	0.01	0.03	0.07	0.1	0.3	0.7	1

www.sematech.org -- 2001 Roadmap

BR 6/00

21

High Performance – Long Term Future Tech Requirements

Table 35b High-performance Logic Technology Requirements—Long-term

YEAR OF PRODUCTION	2010	2013	2016
DRAM 1/2 PITCH (nm)	45	32	22
MPU / ASIC 1/2 PITCH (nm)	50	35	25
MPU PRINTED GATE LENGTH (nm)	25	18	13
MPU PHYSICAL GATE LENGTH (nm)	18	13	9
Physical gate length high-performance (HP) (nm) [1]	18	13	9
Equivalent physical oxide thickness for high-performance T_{ox} (EOT) (nm) [2]	0.5-0.8	0.4-0.6	0.4-0.5
Gate depletion and quantum effects electrical thickness adjustment factor (nm) [3]	0.5	0.5	0.5
T_{ox} electrical equivalent (nm) [4]	1.2	1.0	0.9
Nominal power supply voltage (V_{dd}) (V) [5]	0.6	0.5	0.4

www.sematech.org -- 2001 Roadmap

BR 6/00

22

Design – The Future

Scaling means more transistors.....

Table 18 Additional Design Technology Requirements

YEAR OF PRODUCTION	2001	2002	2003	2004	2005	2006	2007	2010	2013	2016	DRIVER
DRAM ½ PITCH (nm)	130	115	100	90	80	70	65	45	32	22	
MPU / ASIC ½ PITCH (nm)	150	130	107	90	80	70	65	50	35	25	
MPU PRINTED GATE LENGTH (nm)	90	75	65	53	45	40	35	25	18	13	
MPU PHYSICAL GATE LENGTH (nm)	65	53	45	37	32	28	25	18	13	9	
SOC new design cycle (months)	12	12	12	12	12	12	11	11	10	9	SOC
SOC logic Mtx per designer-year (10-person team)	1.2			2.6			6.9	13.6	37.4	117.3	SOC
SOC dynamic power reduction beyond scaling (X)	0			1.5			2.5	4	7	20	SOC
SOC standby power reduction beyond scaling (X)	2			6			16	30	150	800	SOC
%Test covered by BIST	20			30			45	60	75	90	MPU, SOC

Designers need to use more transistors in same time to keep up with increasing transistors, but design challenges grow (i.e, local clock synchronization, clock skew budgets).

BR 6/00

23

MOSFET Scaling (Rabaey text)

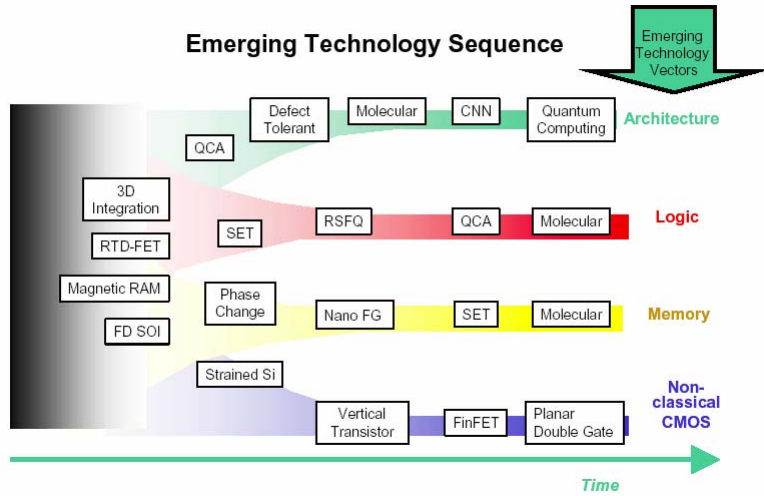
Table 3-9 MOSFET technology projection for high performance logic (from [SIA01]).

Year of Introduction	2001	2003	2005	2007	2010	2013	2016
Drawn channel length (nm)	90	65	45	35	25	18	13
Physical channel length (nm)	65	45	32	25	18	13	9
Gate oxide (nm)	2.3	2.0	1.9	1.4	1.2	1.0	0.9
V_{DD} (V)	1.2	1.0	0.9	0.7	0.6	0.5	0.4
NMOS I_{Dsat} ($\mu A/\mu m$)	900	900	900	900	1200	1500	1500
NMOS I_{leak} ($\mu A/\mu m$)	0.01	0.07	0.3	1	3	7	10

BR 6/00

24

Science Fiction



BR 6/00

25

SF Memory Devices

Table 43 Emerging Research Memory Devices







STORAGE MECHANISM	BASELINE 2002 TECHNOLOGIES		MAGNETIC RAM		PHASE CHANGE MEMORY	NANO FLOATING GATE MEMORY	SINGLE/FEW ELECTRON MEMORIES	MOLECULAR MEMORIES
DEVICE TYPES	DRAM	NOR FLASH	PSEUDO-SPIN-VALVE	MAGNETIC TUNNEL JUNCTION	OUM	-ENGINEERED TUNNEL BARRIER -NANOCRYSTAL	SET	-BISTABLE SWITCH -MOLECULAR NEMS -SPIN BASED MOLECULAR DEVICES
AVAILABILITY	2002		~2004	~2004	~2004	>2005	>2007	>2010
INITIAL F VALUE	130 nm	150 nm	350 nm	130 nm	100 nm	80 nm	65 nm	45 nm

BR 6/00

26

SF Logic Devices

Table 44 Emerging Logic Devices







						
DEVICE	RESONANT TUNNELING DIODE - FET	SINGLE ELECTRON TRANSISTOR	RAPID SINGLE QUANTUM FLUX LOGIC	QUANTUM CELLULAR AUTOMATA	NANOTUBE DEVICES	MOLECULAR DEVICES
TYPES	3-terminal	3-terminal	Josephson Junction + inductance loop	-Electronic QCA -Magnetic QCA	FET	2-terminal and 3-terminal
ADVANTAGES	Density, Performance, RF	Density, Power, Function	High speed, Potentially robust, (insensitive to timing error)	High functional density, No interconnect in signal path, Fast and low power	Density, Power	Identity of individual switches (e.g., size, properties) on sub-nm level. Potential solution to interconnect problem
CHALLENGES	Matching of device properties across wafer	New device and system, Dimensional control (e.g., room temp operation), Noise (offset charge), Lack of drive current	Low temperatures, Fabrication of complex, dense circuitry	Limited fan out, Dimensional control (room temperature operation), Architecture, Feedback from devices, Background charge	New device and system, Difficult route for fabricating complex circuitry	Thermal and environmental stability, Two terminal devices, Need for new architectures
MATURITY	Demonstrated	Demonstrated	Demonstrated	Demonstrated	Demonstrated	Demonstrated

BR 6/00

27

SF Architectures

Table 46 Emerging Research Architectures

						
ARCHITECTURES	3-D INTEGRATION	QUANTUM CELLULAR AUTOMATA	DEFECT TOLERANT ARCHITECTURE	MOLECULAR ARCHITECTURE	CELLULAR NONLINEAR NETWORKS	QUANTUM COMPUTING
DEVICE IMPLEMENTATION	CMOS with dissimilar material systems	Arrays of quantum dots	Intelligently assembles nanodevices	Molecular switches and memories	Single electron array architectures	Spin resonance transistors, NMR devices, Single flux quantum devices
ADVANTAGES	Less interconnect delay, Enables mixed technology solutions	High functional density, No interconnects in signal path	Supports hardware with defect densities >50%	Supports memory based computing	Enables utilization of single electron devices at room temperature	Exponential performance scaling, Enables unbreakable cryptography
CHALLENGES	Heat removal, No design tools, Difficult test and measurement	Limited fan out, Dimensional control (low temperature operation), Sensitive to background charge	Requires pre-computing test	Limited functionality	Subject to background noise, Tight tolerances	Extreme application limitation, Extreme technology
MATURITY	Demonstration	Demonstration	Demonstration	Concept	Demonstration	Concept

BR 6/00

28

Parameters of SF Devices

Table 47 Estimated Parameters for Emerging Research Devices and Technologies in the year 2016

Technology	T _{min} sec	T _{max} sec	CD _{min} m	CD _{max} m	Energy J/op	Cost min \$/gate	Cost max \$/gate
Si CMOS	3E-11 ⁵⁹	1E-6	8E-9	5E-6	4E-18	4E-9	3E-3
RSFQ	1E-12	5E-11	3E-7	1E-6	2E-18	1E-3	1E-2
Molecular	1E-8	1E-3	1E-9	5E-9	1E-20	1E-11	1E-10
Plastic	1E-4	1E-3	1E-4	1E-3	1E-24	1E-9	1E-6
Optical (digital)	1E-16	1E-12	2E-7	2E-6	1E-12	1E-3	1E-2
NEMS (conservative)	1E-7	1E-3	1E-8	1E-7	1E-21	1E-8 ⁶⁰	1E-5
Neuromorphic	1E-13	1E-4	6E-6	6E-6	3E-25	5E-5	1E-2
Quantum	1E-16	1E-15	1E-8	1E-7	1E-21	1E2	1E3

In this table T stands for system cycle time (switching time), CD stands for critical dimension (e.g., physical gate length), Energy is the intrinsic operational energy of one device, and Cost is defined as \$ per gate.

Summary

Get ready to re-invent yourself around 2016!!