CMOS Technology Scaling

- See page 146-149 of Rabaey text
- Full Scaling – Vdd scaled by S, Dimensions by S
- General Scaling - Vdd 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)

Long Channel Scaling, Delay

Drain Current, MOS transistor, in Saturation, Long Channel

$I_{DSAT} = K'_n/2 * W/L *(V_{GS} - V_T)^2 (1 + \lambda * V_{DS})$ pg. 45

Recall that $K'_n = \mu_n Cox$, so will scale by S because Cox increases with decreasing thickness. If K'_n scales by S, then so will I_D.

C_L is gate capacitance = Cox * W *L
Scales by 1/S because both W, L scale by 1/S but Cox scales by S (Cox increases with decreasing thickness).

Intrinsic Delay = $C_L * V / I_D$; so will scale by $1/S^2$!!

Short Channel, Velocity Saturation

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

$$\nu_n = -\mu_n E(x) = \mu_n \, dV/dx$$ pg 44.

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 Esat, the velocity of the carriers saturate (Figure 2.28, pg 53, textbook).

For p-type silicon (NMOS transistor), Esat = 1.5V μm

Easily reached with channel lengths < 1.0 μm .

Short Channel I_{DSAT} Equation

A new equation for IDSAT

$$IDSAT = \nu_{SAT} \, C_{ox} \, W \, (V_{GS} - V_{DSAT} - V_T)$$ pg 54.

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.

Short Channel Scaling, Delay

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 (see Fig 2.28, pg 53). This will also decrease the delay scaling for short channel devices.

Power Scaling

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

where Tp is intrinsic delay.

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

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

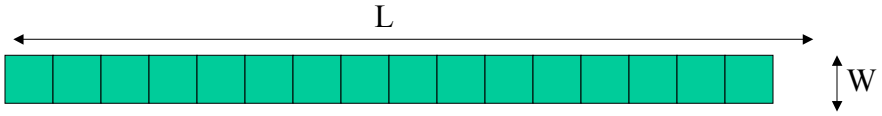
What about power density??? (power per unit area). Even with Pav being '1' for short channel devices, if we had N devices before in a given area, we can now pack N2 devices in the same area since W, L are both scaled by 1/S.

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

Wire Resistance Scaling

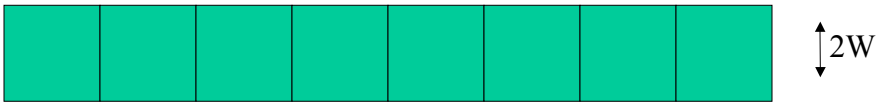
Wire Resistance = ohms/square * L / W where

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



R wire = R_{sq} * L / W

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



R wire_new = R_{sq} * L / 2W = R wire old/2

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

Rwire is constant for local wiring (both L, W decrease).

Rwire scales by S by global wiring since W decreases but L remains the same.

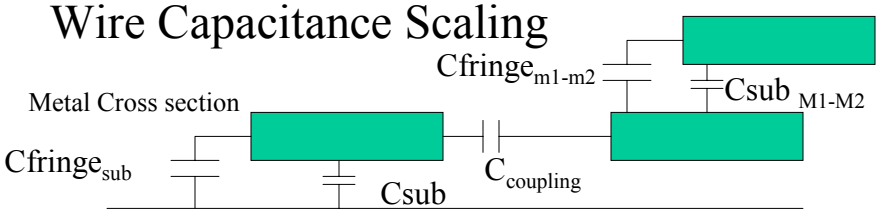
Sheet Resistance (Leda 0.25U)

	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 e-8 Ohm-meters
Copper Resistivity: 1.67 e-8 Ohm-meters

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

Wire Capacitance Scaling



Cwire = Cfringe + Csub + Ccoupling

Csub = Cox* W * L

Recall that Cox scales by S. So Csub scales by 1/S.

Cfringe depends on thickness of sidewall, and L of wire, Cox of insulator. Thickness will remain constant.

Cfringe (sub) will be constant (L scales by 1/S, Cox by S). Ccoupling is *controlled via spacing rules*. In sub-micron technologies, minimum spacing often controlled by capacitance considerations.

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.

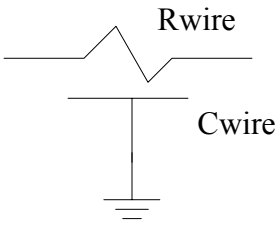
Interconnect Capacitance (Fringe)

	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.

Wire Delay



If Csub dominates, the Cwire scales by 1/S
Rwire is constant for local wires, so local wire Rwire*Cwire (delay) scales by 1/S which is good news (gate delay also scales by 1/S).

Rwire scales by S for global wires, so global wire delay Rwire*Cwire(delay) is constant!!! The gate delays scale down, so global wire delay scales UPWARD with respect to gate delays. BAD!!

Clock Speed Scaling

Most systems have less than 16 FO4 delays between registers.

System clock speed determined by:
Clock2Q of Register + Regster2Register Delay + Setup + clock Skew budget

Clock2Q, Setup, Register2Register delay scales down with techology.

However, clock is a global signal. Clock skews remain constant, and grow releative 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.

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.

IJSSC May '98 "High-performance Microprocessor Design

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.

Vdd - The Future

Table 6a Power Supply and Power Dissipation—Near Term Years

YEAR / *TECHNOLOGY NODE*	*1999* *180 nm*	*2000*	*2001*	*2002* *130 nm*	*2003*	*2004*	*2005* *100 nm*
Power Supply Voltage (V)							
Minimum logic V_{dd} (V)—maximum (for maximum performance)	1.8	1.8	1.5	1.5	1.5	1.2	1.2
Minimum logic V_{dd} (V)—minimum (for lowest power))	1.5	1.5	1.2	1.2	1.2	0.9	0.9
Maximum Power							
High-performance with heatsink (W)	90	100	115	130	140	150	160
Battery (W)—(hand-held)	1.4	1.6	1.7	2.0	2.1	2.3	2.4

www.sematech.org -- 1999 Roadmap

Note two scenarios -- Maximum performance or lowest power.

Vdd - The Future (cont)

Table 6b Power Supply and Power Dissipation—Long Term Years

YEAR / *TECHNOLOGY NODE*	*2008* *70 nm*	*2011* *50 nm*	*2014* *35 nm*
Power Supply Voltage (V)			
Minimum logic V_{dd} (V)—maximum (for maximum performance)	0.9	0.6	0.60
Minimum logic V_{dd} (V)—minimum (for lowest power))	0.6	0.5	0.30
Maximum Power			
High-performance with heatsink (W)	170	174	183
Battery (W)—(hand-held)	2.0	2.2	2.4

Design – The Future

Scaling means more transistors.....

Table 14a Design Technology Requirements—Near Term

YEAR / *TECHNOLOGY NODE*	*1999* *180 nm*	*2000*	*2001*	*2002* *130 nm*	*2003*	*2004*	*2005* *100 nm*
MPU new design cycle (months)	36	36	36	32	32	32	30
MPU transistors per designer-month (300-person team) (thousand)	2	3	4	7	10	15	20
ASIC new design cycle (months)	12	12	12	12	12	12	12
ASIC transistors per designer-month (50-person team) (million)	0.3	0.4	0.5	0.7	1.0	1.3	1.8
Portion of verification by formal methods	15%	15%	15%	20%	20%	20%	30%
Portion of test covered by BIST	20%	20%	20%	30%	30%	30%	40%

Requirements on designers (Sematech 1999 Roadmap).

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).