THE INVERTER DYNAMICS

[Adapted from Rabaey's *Digital Integrated Circuits*, ©2002, J. Rabaey et al.]

Dynamic Behavior

Propagation Delay, T_p

- Defines how quickly output is affected by input
- Measured between 50% transition from input to output
- t_{pLH} defines delay for output going from low to high
- t_{pHL} defines delay for output going from high to low
- Overall delay, t_p, defined as the of t_{pLH} and t_{pHL}

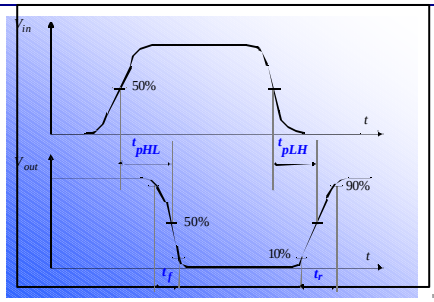
Dynamic Behavior

Rise and fall time, T_r and T_f

- Defines slope of the signal
- Defined between the 10% and 90% of the signal swing

Propagation delay and rise and fall times affected by the fan-out due to larger capacitance loads

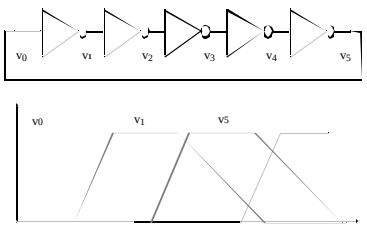
Delay Definitions



The Ring Oscillator

- A standard method is needed to measure the gate delay
- It is based on the ***ring oscillator***
- $2Nt_p >> t_f + t_r$ for proper operation

Ring Oscillator



$T = 2 \times t_p \times N$

Power Dissipation

- Power consumption determines heat dissipation and energy consumption
- Power influences design decisions:
  - ***packaging and cooling***
  - width of supply lines
  - power-supply capacity
  - # of transistors integrated on a single chip

Power requirements make high density bipolar ICs impossible (feasibility, cost, reliability)

Power Dissipation

Supply-line sizing

$$P_{peak} = i_{peak} V_{supply} = max(p(t))$$

$$P_{av} = \frac{1}{T}\int_0^T p(t)dt = \frac{V_{supply}}{T}\int_0^T i_{supply}(t)dt$$

Battery drain, cooling

Power Dissipation

- P_{peak} = static power + dynamic power
- Dynamic power:
  - (dis)charging capacitors
  - temporary paths from VDD to VSS
  - proportional to switching frequency
- Static power:
  - static conductive paths between rails
  - leakage
  - increases with temperature

Power Dissipation

- Propagation delay is related to power consumption
  - t_p determined by speed of charge transfer
  - fast charge transfer => fast gate
  - fast gate => more power consumption
- *Power-delay product (PDP)*
  - quality measure for switching device
  - PDP = energy consumed / gate / switching event
  - measured using ring oscillator

Power Dissipation

Supply-line sizing

$$P_{peak} = i_{peak} V_{supply} = max(p(t))$$

$$P_{av} = \frac{1}{T}\int_0^T p(t)dt = \frac{V_{supply}}{T}\int_0^T i_{supply}(t)dt$$

Battery drain, cooling

Energy consumed / gate / switching event

Power-Delay Product

$$PDP = t_p \times P_{av}$$

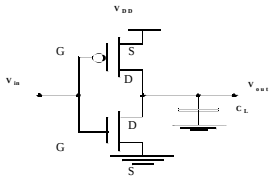
= Energy dissipated per operation

CMOS Inverter: Steady State Response

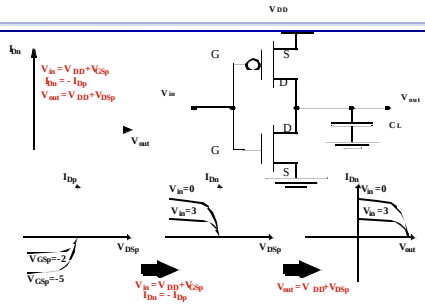
- **CMOS technology:**
  - **No path exists between VDD and VSS in steady state**
  - **No static power consumption! (ideally)**
  - **Main reason why CMOS replaced NMOS in early 80's**
- **NMOS technology:**
  - **Has NMOS pull-up device that is always ON**
  - **Creates voltage divider when pull-down is ON**
  - **Power consumption puts upper bound on (# devices / chip)**

Voltage Transfer Characteristic

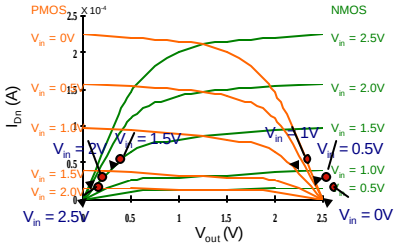
CMOS Inverter Load Characteristics



PMOS Load Lines

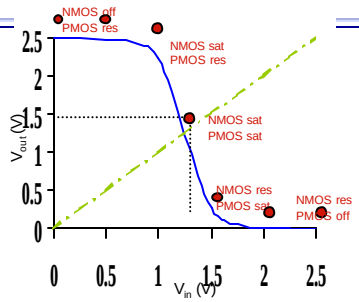


CMOS Inverter Load Lines



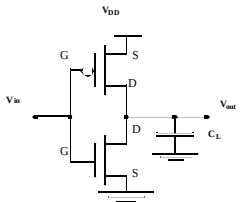
0.25um, W/L_n = 1.5, W/L_p = 4.5, V_{DD} = 2.5V, V_{Tn} = 0.4V, V_{tp} = -0.4V

CMOS Inverter VTC



	Cutoff	Linear	Saturation
pMOS	$V_{in} - V_{DD} = V_{GS} > V_T$	$V_{in} - V_{DD} = V_{GS} < V_T$ $V_{in} - V_{out} = V_{GD} < V_T$	$V_{in} - V_{DD} = V_{GS} > V_T$ $V_{in} - V_{out} = V_{GD} > V_T$
nMOS	$V_{in} = V_{GS} < V_T$	$V_{in} = V_{GS} > V_T$ $V_{in} - V_{out} = V_{GD} > V_T$	$V_{in} = V_{GS} > V_T$ $V_{in} - V_{out} = V_{GD} < V_T$

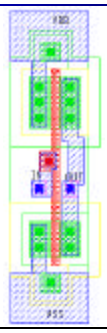
Regions of operations
For nMOS and pMOS
In CMOS inverter



CMOS Inverter Load Characteristics

- **For valid dc operating points:**
  - **current through NMOS = current through PMOS**
  - **=> dc operating points are the intersection of load lines**
- **All operating points located at high or low output levels**
  - **=> VTC has narrow transition zone**
  - **high gain of transistors during switching**
  - **transistors in saturation**
    - **high transconductance (g_m)**
    - **high output resistance (voltage controlled current source)**

Voltage Transfer Characteristic



Switching Threshold

- V_M where $V_{in} = V_{out}$ (both PMOS and NMOS in saturation since $V_{DS} = V_{GS}$)

  $V_M \approx rV_{DD}/(1 + r)$ where $r = k_pV_{DSATp}/k_nV_{DSATn}$
- Switching threshold set by the ratio r, which compares the relative driving strengths of the PMOS and NMOS transistors
- Want $V_M = V_{DD}/2$ (to have comparable high and low noise margins), so want $r \approx 1$

$$\frac{(W/L)_p}{(W/L)_n} = \frac{k_n'V_{DSATn}(V_M - V_{Tn} - V_{DSATn}/2)}{k_p'V_{DSATp}(V_{DD}-V_M+V_{Tp}+V_{DSATp}/2)}$$

Switch Threshold Example

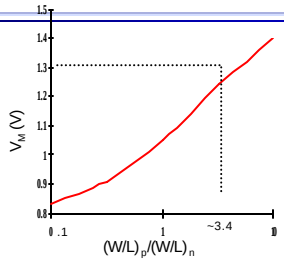
- In 0.25 µm CMOS process, using parameters from table, V_{DD} = 2.5V, and minimum size NMOS ($(W/L)_n$ of 1.5)

	V_{T0}(V)	$\gamma(V^{0.5})$	V_{DSAT}(V)	$k'(AN/V^2)$	$\lambda(V^{-1})$
NMOS	0.43	0.4	0.63	115 x 10^{-6}	0.06
PMOS	-0.4	-0.4	-1	-30 x 10^{-6}	-0.1

$$\frac{(W/L)_p}{(W/L)_n} = \frac{115 \times 10^{-6}}{-30 \times 10^{-6}} \times \frac{0.63}{-1.0} \times \frac{(1.25 - 0.43 - 0.63/2)}{(1.25 - 0.4 - 1.0/2)} = 3.5$$

$(W/L)_p$ = 3.5 x 1.5 = 5.25 for a V_M of 1.25V

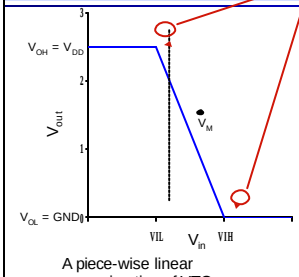
Simulated Inverter V_M



- V_M is relatively insensitive to variations in device ratio
  - setting the ratio to 3, 2.5 and 2 gives V_M's of 1.22V, 1.18V, and 1.13V
- Increasing the width of the PMOS moves V_M towards V_{DD}

Note: **x-axis is semilog**

Noise Margins Determining V_{IH} and V_{IL}



By definition, V_{IH} and V_{IL} are where $dV_{out}/dV_{in} = -1$ (= gain)

$NM_H = V_{DD} - V_{IH}$

$NM_L = V_{IL} - GND$

Approximating:

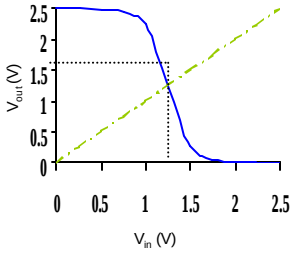
$V_{IH} = V_M - V_M/g$

$V_{IL} = V_M + (V_{DD} - V_M)/g$

A piece-wise linear approximation of VTC

So high gain in the transition region is very desirable

CMOS Inverter VTC from Simulation



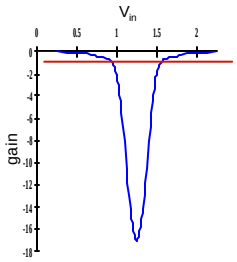
0.25um, $(W/L)_p/(W/L)_n = 3.4$
$(W/L)_n = 1.5$ (min size)
$V_{DD} = 2.5V$

$V_M \approx 1.25V$, $g = -27.5$

$V_{IL} = 1.2V$, $V_{IH} = 1.3V$
$NM_L = NM_H = 1.2$
(actual values are
$V_{IL} = 1.03V$, $V_{IH} = 1.45V$
$NM_L = 1.03V$ & $NM_H = 1.05V$)

Output resistance
low-output = 2.4kΩ
high-output = 3.3kΩ

Gain Determinates



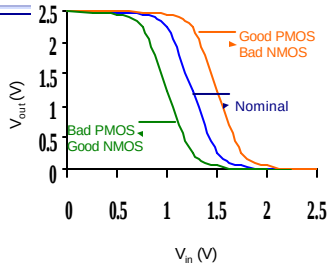
Gain is a strong function of the slopes of the currents in the saturation region, for $V_{in} = V_M$

$$g \approx \frac{(1+r)}{(V_M - V_{Tn} - V_{DSATn}/2)(\lambda_n - \lambda_p)}$$

✓Determined by technology parameters, especially λ.

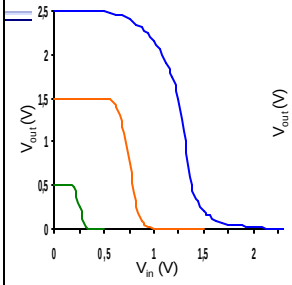
✓Only designer influence through supply voltage and V_M (transistor sizing).

Impact of Process Variation

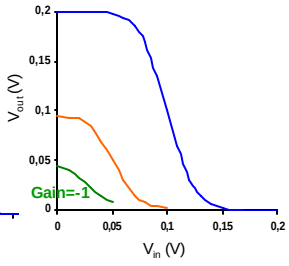


process variations (mostly) cause a shift in the switching threshold

Scaling the Supply Voltage



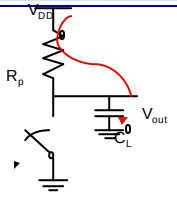
Device threshold voltages are kept (virtually) constant



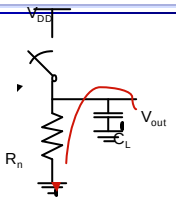
Device threshold voltages are kept (virtually) constant

Propagation Delay

Switch Model of Dynamic Behavior



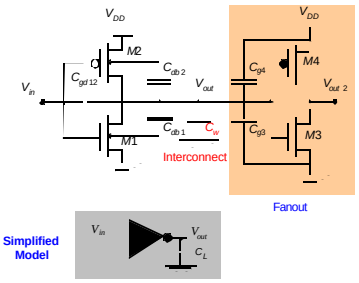
$V_{in} = 0$



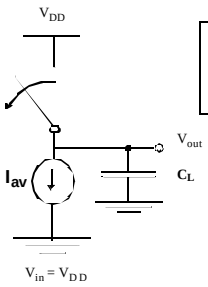
$V_{in} = V_{DD}$

Gate response time is determined by the time to charge C_L through R_p (discharge C_L through R_n)

What is the Inverter Driving?



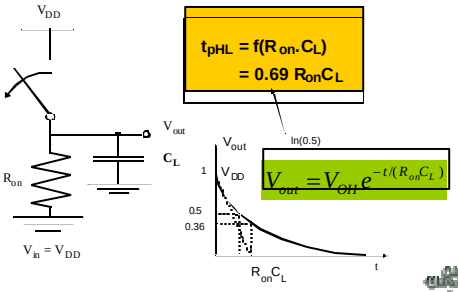
CMOS Inverter Propagation Delay Approach 1



$$t_{pHL} = \frac{C_L V_{swing}/2}{I_{av}}$$

$$\sim \frac{C_L}{k_n V_{DD}}$$

CMOS Inverter Propagation Delay Approach 2

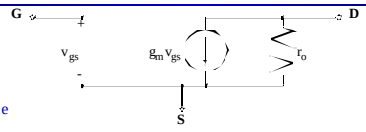


CMOS Inverter: Transient Response

How can the designer build a fast gate?

- $t_{pHL} = f(R_{on} * C_L)$
- **Keep output capacitance, C_L, small**
  - **low fan-out**
  - **keep interconnections short (floor-plan your layout!)**
- **Decrease on-resistance of transistor**
  - **increase W/L ratio**
  - **make good contacts (slight effect)**

MOS Transistor Small Signal Model



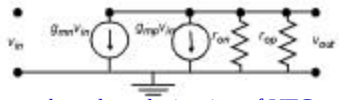
Define

	g_m	r_o
linear	kV'_{DS}	$[k(V_{GS}-V_T-V'_{DS})]^{-1}$
saturation	$k(V_{GS}-V_T)$	$1/\lambda I_D$

Determining V_{IH} and V_{IL}

At V_{IH} (V_{IL}): $\frac{\delta V_{out}}{\delta V_{in}} = -1$

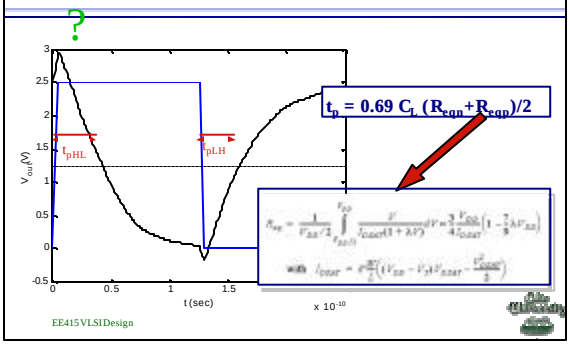
small-signal model of inverter



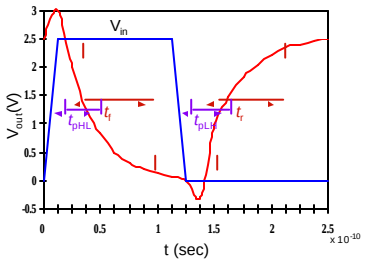
V_{IH} and V_{IL} are based on derivative of VTC equal to -1

$$g = \frac{v_{out}}{v_{in}} = -(g_{mn} + g_{mp}) \times (r_{on} \| r_{op}) = -1$$

Transient Response



Inverter Transient Response

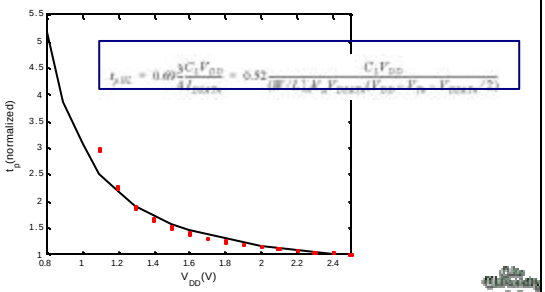


V_{DD}=2.5V
0.25µm
W/L_n = 1.5
W/L_p = 4.5
R_{eqn}= 13 kΩ (÷ 1.5)
R_{eqp}= 31 kΩ (÷ 4.5)

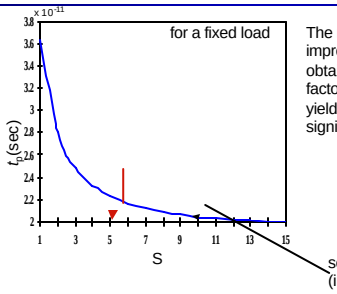
t_{pHL} = 36 psec
t_{pLH} = 29 psec
so
t_p = 32.5 psec

From simulation: t_{pHL} = 39.9 psec and t_{pLH} = 31.7 psec

Delay as a function of V_{DD}



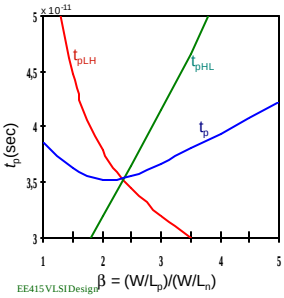
Sizing Impacts on Delay



The majority of the improvement is already obtained for S = 5. Sizing factors larger than 10 barely yield any extra gain (and cost significantly more area).

self-loading effect (intrinsic capacitance dominates)

PMOS / NMOS Ratio Effects

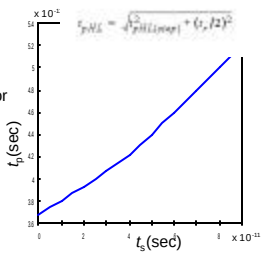


β of 2.4 (= 31 kΩ/13 kΩ) gives symmetrical response

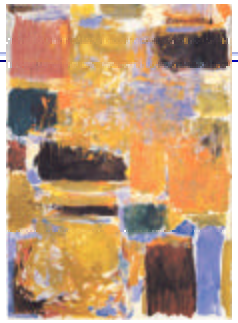
β of 1.6 to 1.9 gives optimal performance

Input Signal Rise / Fall Time

- In reality, the input signal changes gradually (and both PMOS and NMOS conduct for a brief time). This affects the current available for charging/discharging C_L and impacts propagation delay.
- t_p increases linearly with increasing input rise time, t_r, once $t_r > t_p$
- t_r is due to the limited driving capability of the preceding gate

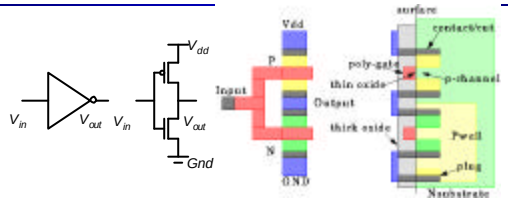


for a minimum-size inverter with a fan-out of a single gate



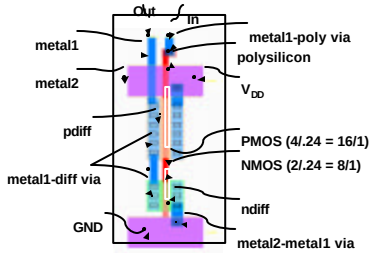
Inverter Sizing

CMOS Inverter: Four Views



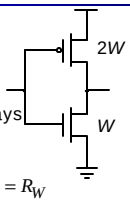
Logic Transistor Layout Physical

CMOS Inverter Sizing



Inverter Delay

- Minimum length devices, L=0.25µm
- Assume that for $W_P = 2W_N = 2W$
  - same pull-up and pull-down currents
  - approx. equal resistances $R_N = R_P$
  - approx. equal rise t_{pLH} and fall t_{pHL} delays
- Analyze as an RC network

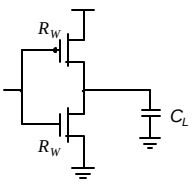


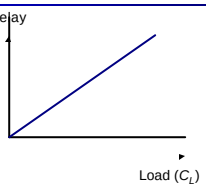
$$R_P = R_{unit}\left(\frac{W_P}{W_{unit}}\right)^{-1} \approx R_{unit}\left(\frac{W_N}{W_{unit}}\right)^{-1} = R_N = R_W$$

Delay (D): $t_{pHL} = (\ln 2)\, R_N C_L$ $\quad t_{pLH} = (\ln 2)\, R_P C_L$

Load for the next stage: $C_{gin} = 3\dfrac{W}{W_{unit}} C_{unit}$

Inverter with Load





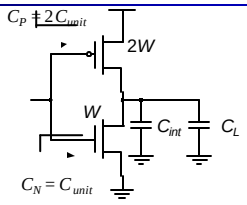
$t_p = k\, R_W C_L$

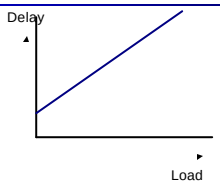
k is a constant, equal to 0.69

Assumptions: no load -> zero delay

W_{unit} = 1

Inverter with Load





Delay = $kR_W(C_{int} + C_L) = kR_W C_{int} + kR_W C_L = kR_W C_{int}(1 + C_L / C_{int})$
= Delay (Internal) + Delay (Load)

Delay Formula

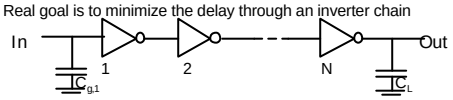
$$Delay \sim R_W (C_{int} + C_L)$$

$$t_p = kR_W C_{int}(1 + C_L / C_{int}) = t_{p0}(1 + f/g)$$

$C_{int} = gC_{gin}$ with $g \gg 1$
$f = C_L/C_{gin}$ - effective fanout
$R = R_{unit}/W$; $C_{int} = WC_{unit}$
$t_{p0} = 0.69 R_{unit} C_{unit}$

Inverter Chain

- Real goal is to minimize the delay through an inverter chain



the delay of the j-th inverter stage is

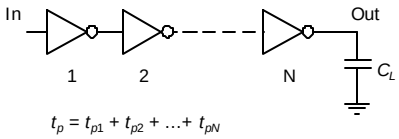
$$t_{p,j} = t_{p0}(1 + C_{g,j+1}/(\gamma C_{g,j})) = t_{p0}(1 + f_j/\gamma)$$

and $t_p = t_{p1} + t_{p2} + \ldots + t_{pN}$

so $t_p = \Sigma t_{p,j} = t_{p0} \Sigma (1 + C_{g,j+1}/(\gamma C_{g,j}))$

- If C_L is given
  - » How should the inverters be sized?
  - » How many stages are needed to minimize the delay?

Apply to Inverter Chain



$$t_p = t_{p1} + t_{p2} + \ldots + t_{pN}$$

$$t_{pj} \sim R_{unit} C_{unit}\left(1 + \frac{C_{gin,j+1}}{gC_{gin,j}}\right)$$

$$t_p = \sum_{j=1}^{N} t_{p,j} = t_{p0} \sum_{i=1}^{N}\left(1 + \frac{C_{gin,j+1}}{gC_{gin,j}}\right),\ C_{gin,N+1} = C_L$$

Optimum Delay and Number of Stages

When each stage is sized by f and has same eff. fanout f:

$$f^N = F = C_L / C_{gin,1}$$

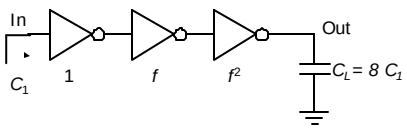
Effective fanout of each stage:

$$f = \sqrt[N]{F}$$

Minimum path delay

$$t_p = N t_{p0}\left(1 + \sqrt[N]{F}/g\right)$$

Example



C_L/C_1 has to be evenly distributed across $N = 3$ stages:

$$f = \sqrt[3]{8} = 2$$

Optimal Number of Inverters

- What is the optimal value for N given F (=f^N) ?
  - » if the number of stages is too large, the intrinsic delay dominates
  - » if the number of stages is too small, the effective fan-out dominates
- The optimum N is found by differentiating the minimum delay divided by the number of stages and setting the result to 0,
- For $\gamma = 0$ (ignoring self-loading) N = ln (F) and the effective fan out becomes f = e = 2.71828

Optimum Number of Stages

For a given load, C_L and given input capacitance C_{in}
Find optimal sizing f

$$C_L = F \cdot C_{in} = f^N C_{in} \quad \text{with} \quad N = \frac{\ln F}{\ln f}$$

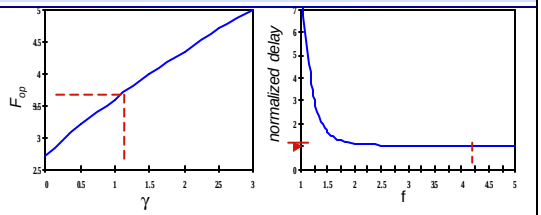
$$t_p = N t_{p0}\left(F^{1/N}/g + 1\right) = \frac{t_{p0} \ln F}{g}\left(\frac{f}{\ln f} + \frac{g}{\ln f}\right)$$

$$\frac{\partial t_p}{\partial f} = \frac{t_{p0} \ln F}{g} \cdot \frac{\ln f - 1 - g/f}{\ln^2 f} = 0$$

For g = 0, f = e, N = ln$F$

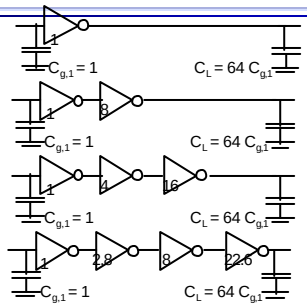
$$f = \exp(1 + g/f)$$

Optimum Effective Fan-Out



- Choosing f larger than optimum has little effect on delay and reduces the number of stages (and area).
  - » Common practice to use f = 4 (for γ = 1)
  - » But too many stages has a substantial negative impact on delay

Example of Inverter (Buffer) Staging



N	f	t_p
1	64	65
2	8	18
3	4	15
4	2.8	15.3

Impact of Buffer Staging for Large C_L

F (g = 1)	Unbuffered	Two Stage Chain	Opt. Inverter Chain
10	11	8.3	8.3
100	101	22	16.5
1.000	1001	65	24.8
10.000	10.001	202	33.1

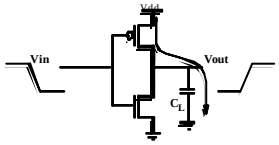
- Impressive speed-ups with optimized cascaded inverter chain for very large capacitive loads.

Design Challenge

- Keep signal rise times smaller than or equal to the gate propagation delays.
  - » good for performance
  - » good for power consumption
- Keeping rise and fall times of the signals small and of approximately equal values is one of the major challenges in high-performance designs - slope engineering.

Power Dissipation

Dynamic Power Dissipation

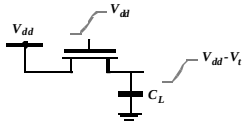


Energy/transition = $C_L * V_{dd}^2$

Power = Energy/transition * $f = C_L * V_{dd}^2 * f$

- Not a function of transistor sizes!
- Need to reduce C_L, V_{dd}, and f to reduce power.

Modification for Circuits with Reduced Swing



$$E_{0\rightarrow 1} = C_L \bullet V_{dd} \bullet (V_{dd} - V_t)$$

- Can exploit reduced swing to lower power (e.g., reduced bit-line swing in memory)

Node Transition Activity and Power

- Consider switching a CMOS gate for N clock cycles

$$E_N = C_L \bullet V_{dd}^2 \bullet n(N)$$

E_N : the energy consumed for N clock cycles

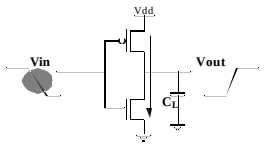
$n(N)$: the number of 0->1 transition in N clock cycles

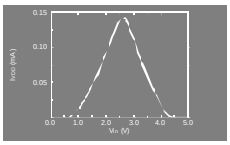
$$P_{avg} = \lim_{N\rightarrow\infty} \frac{E_N}{N} \bullet f_{clk} = \left(\lim_{N\rightarrow\infty} \frac{n(N)}{N}\right) \bullet C_L \bullet V_{dd}^2 \bullet f_{clk}$$

$$\alpha_{0\rightarrow 1} = \lim_{N\rightarrow\infty} \frac{n(N)}{N}$$

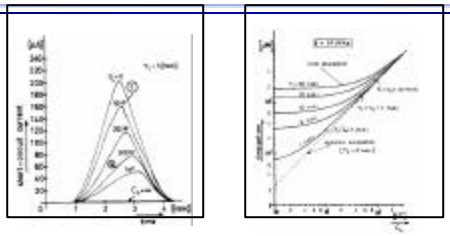
$$P_{avg} = \alpha_{0\rightarrow 1} \bullet C_L \bullet V_{dd}^2 \bullet f_{clk}$$

Short Circuit Currents



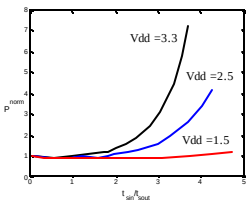


How to keep Short-Circuit Currents Low?



Short circuit current goes to zero if $t_{fall} >> t_{rise}$, but can't do this for cascade logic, so ...

Minimizing Short-Circuit Power



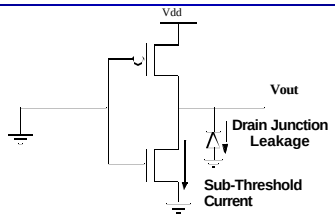
- Keep the input and output rise/fall times the same (< 10% of Total Consumption)

from [Veendrick84]

(IEEE Journal of Solid-State Circuits, August 1984)

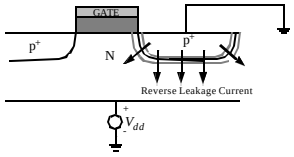
- If $V_{dd} < V_{tn} + |V_{tp}|$ then short-circuit power can be *eliminated*!

Leakage



Sub-threshold current one of most compelling issues in low-energy circuit design!

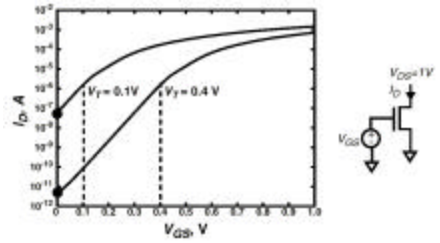
Reverse-Biased Diode Leakage



$I_{DL} = J_S \times A$

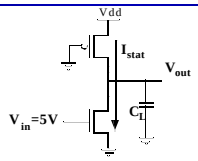
JS = 10-100 pA/μm2 at 25 deg C for 0.25μm CMOS
JS doubles for every 9 deg C!

Subthreshold Leakage Component



- Leakage control is critical for low-voltage operation

Static Power Consumption

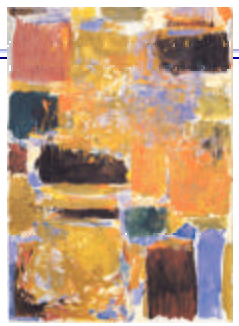


$P_{stat} = P_{(In=1)} . V_{dd} . I_{stat}$

Wasted energy …
Should be avoided in almost all cases,
but could help reducing energy in others (e.g. sense amps)

Principles for Power Reduction

- Prime choice: Reduce voltage!
  - » Recent years have seen an acceleration in supply voltage reduction
  - » Design at very low voltages still open question (0.6 … 0.9 V by 2010!)
- Reduce switching activity
- Reduce physical capacitance
  - » Device Sizing: for *F*=20
    - – f_{opt}(energy)=3.53, f_{opt}(performance)=4.47



Impact of Technology Scaling

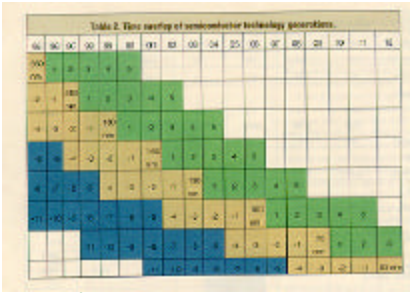
Goals of Technology Scaling

- Make things cheaper:
  - » Want to sell more functions (transistors) per chip for the same money
  - » Build same products cheaper, sell the same part for less money
  - » Price of a transistor has to be reduced
- But also want to be faster, smaller, lower power

Technology Scaling

- Goals of scaling the dimensions by 30%:
  - » Reduce gate delay by 30% (increase operating frequency by 43%)
  - » Double transistor density
  - » Reduce energy per transition by 65% (50% power savings @ 43% increase in frequency
- Die size used to increase by 14% per generation
- Technology generation spans 2-3 years

Technology Generations



Technology Evolution (2000 data)

International Technology Roadmap for Semiconductors

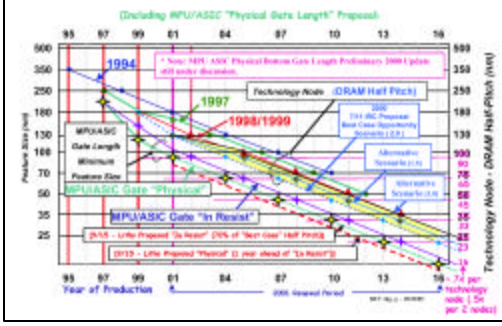
Year of Introduction	1999	2000	2001	2004	2008	2011	2014
Technology node [nm]	180		130	90	60	40	*30*
Supply [V]	1.5-1.8	1.5-1.8	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6	*0.3-0.6*
Wiring levels	6-7	6-7	7	8	9	9-10	*10*
Max frequency [GHz],Local-Global	1.2	1.6-1.4	2.1-1.6	3.5-2	7.1-2.5	11-3	*14.9 3.6*
Max μP power [W]	90	106	130	160	171	177	*186*
Bat. power [W]	1.4	1.7	2.0	2.4	2.1	2.3	*2.5*

Node years: 2007/65nm, 2010/45nm, 2013/33nm, 2016/23nm

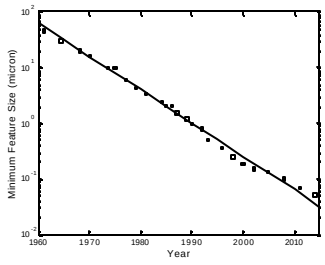
Technology Evolution (1999)

Year of Introduction	1994	1997	2000	2003	2006	2009
Channel length (μm)	0.4	0.3	0.25	0.18	0.13	0.1
Gate oxide (nm)	12	7	6	4.5	4	4
V_{DD} (V)	3.3	2.2	2.2	1.5	1.5	1.5
V_T (V)	0.7	0.7	0.7	0.6	0.6	0.6
NMOS I_{Dsat} (mA/μm) (@ $V_{GS} = V_{DD}$)	0.35	0.27	0.31	0.21	0.29	0.33
PMOS I_{Dsat} (mA/μm) (@ $V_{GS} = V_{DD}$)	0.16	0.11	0.14	0.09	0.13	0.16

ITRS Technology Roadmap Acceleration Continues

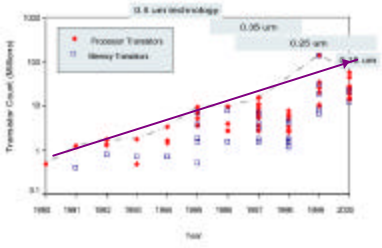


Technology Scaling (1)



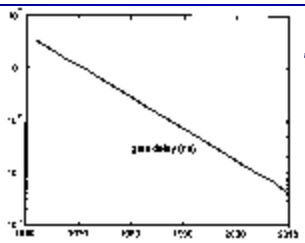
Minimum Feature Size

Technology Scaling (2)



Number of components per chip

Technology Scaling (3)



t_p decreases by 13%/year
50% every 5 years!

Propagation Delay

Technology Scaling Models

- **Full Scaling (Constant Electrical Field)**
  - ideal model — dimensions and voltage scale together by the same factor *S*
- **Fixed Voltage Scaling**
  - most common model until recently — only dimensions scale, voltages remain constant
- **General Scaling**
  - most realistic for todays situation — voltages and dimensions scale with different factors

Scaling Relationships for Long Channel Devices

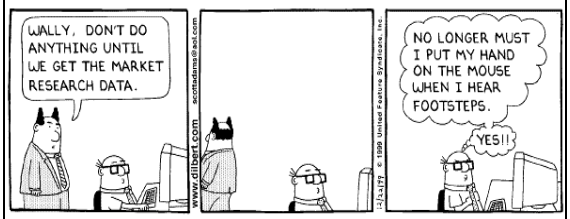
Parameter	Relation	Full Scaling	General Scaling	Fixed Voltage Scaling
W, L, t_{ox}		1/S	1/S	1/S
V_{DD}, V_T		1/S	1/U	1
N_{SUB}	V/W_{dep}^2	S	S^2/U	S^2
Area/Device	WL	$1/S^2$	$1/S^2$	$1/S^2$
C_{ox}	$1/t_{ox}$	S	S	S
C_L	$C_{ox}WL$	1/S	1/S	1/S
k_n, k_p	$C_{ox}W/L$	S	S	S
I_{ds}	$k_{n,p}V^2$	1/S	S/U^2	S
t_p (intrinsic)	C_LV / I_{ds}	1/S	U/S^2	$1/S^2$
P_{ds}	C_LV^2/t_p	$1/S^2$	S/U^3	S
PDP	C_LV^2	$1/S^3$	$1/SU^2$	1/S

Transistor Scaling (velocity-saturated devices)

Parameter	Relation	Full Scaling	General Scaling	Fixed Voltage Scaling

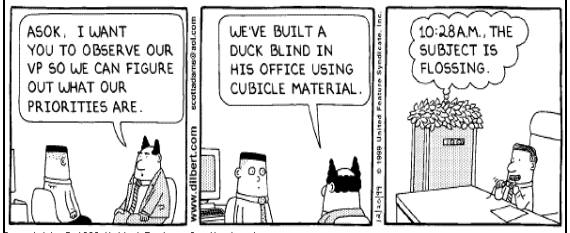
Dilbert

Dilbert



Dilbert

Dilbert



Dilbert