数字集成电路

第五讲 反相器

5.1 简介

- 反相器是所有数字设计的核心
  - 一旦清楚了反相器的工作原理和性质，设计其它逻辑门和复杂逻辑（加法器、乘法器和微处理器等）就大大简化了。
  - 复杂电路的电气特性几乎完全可以由反相器中得到的结果推导出来。
  - 反相器的分析可以延伸来解释比较复杂的门（如**NAND**、**NOR**）的特性。

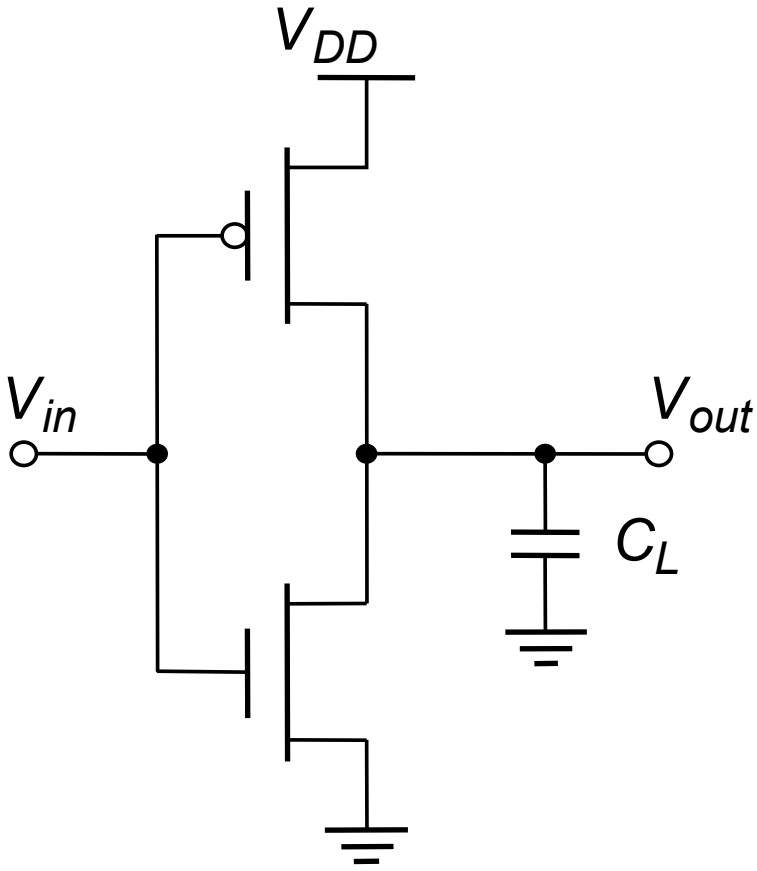
简介

- CMOS反相器的分析：
  - 成本：用复杂性和面积表示
  - 完整性和稳定性：用静态（稳态）特性表示
  - 性能：由动态（即瞬态）响应决定
  - 能耗效率：由能耗和功耗决定
- 按比例缩小技术的影响
- 反相器是分析组合逻辑和时序逻辑的基础

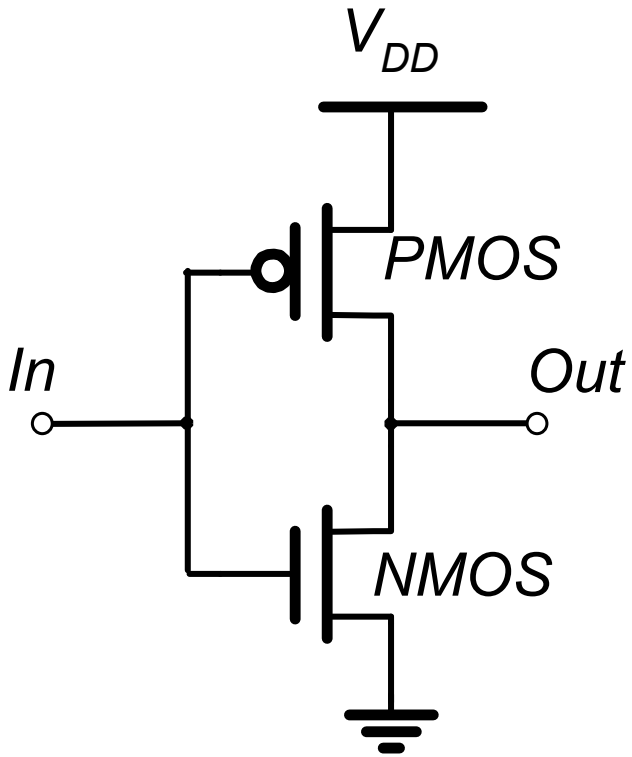
内容提要

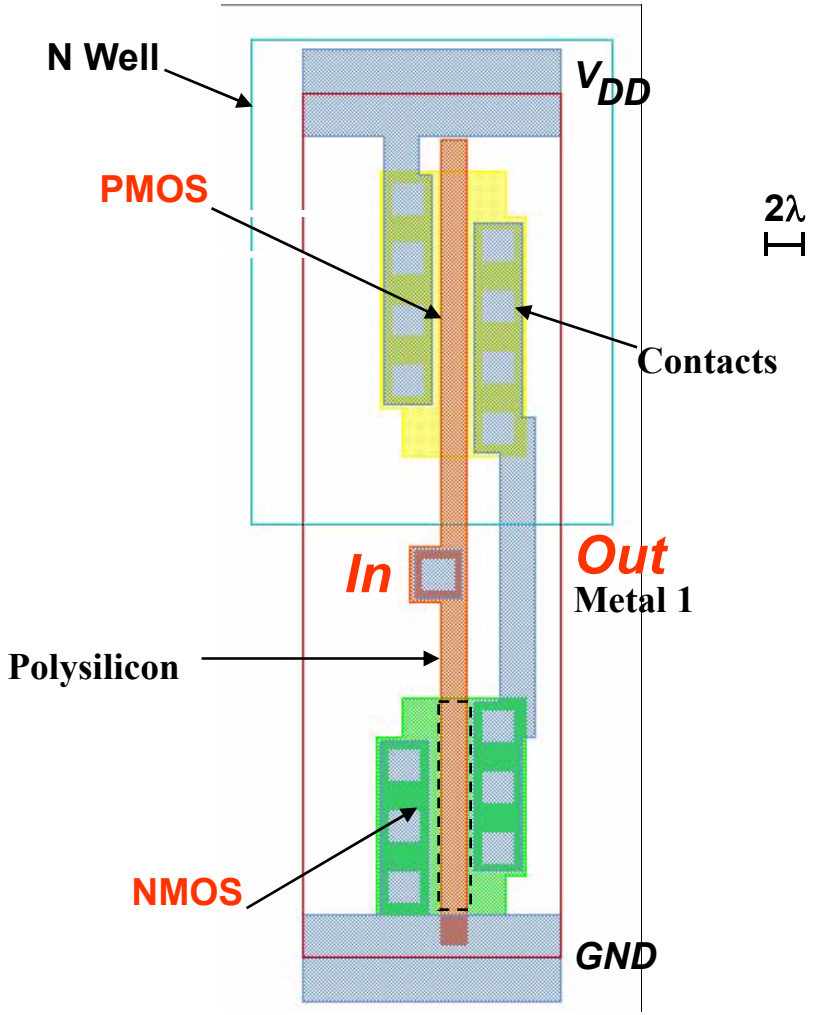
- 直观综述
- 电压传输特性（VTC）
- 可靠性：静态特性
- 性能：动态特性
- 功耗和能耗—延时积
- 按比例缩小技术以及对反相器的影响

5.2 The CMOS Inverter: A First Glance



CMOS Inverter

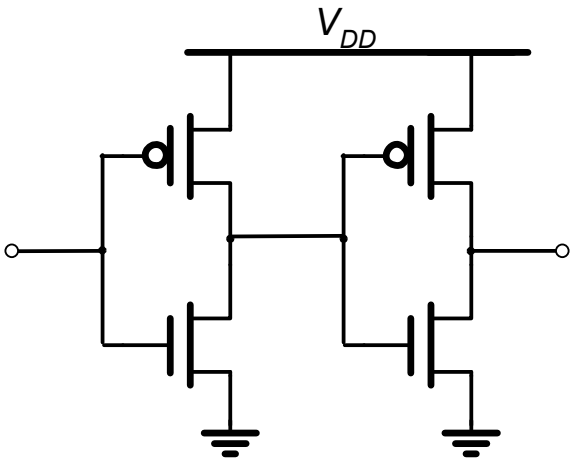


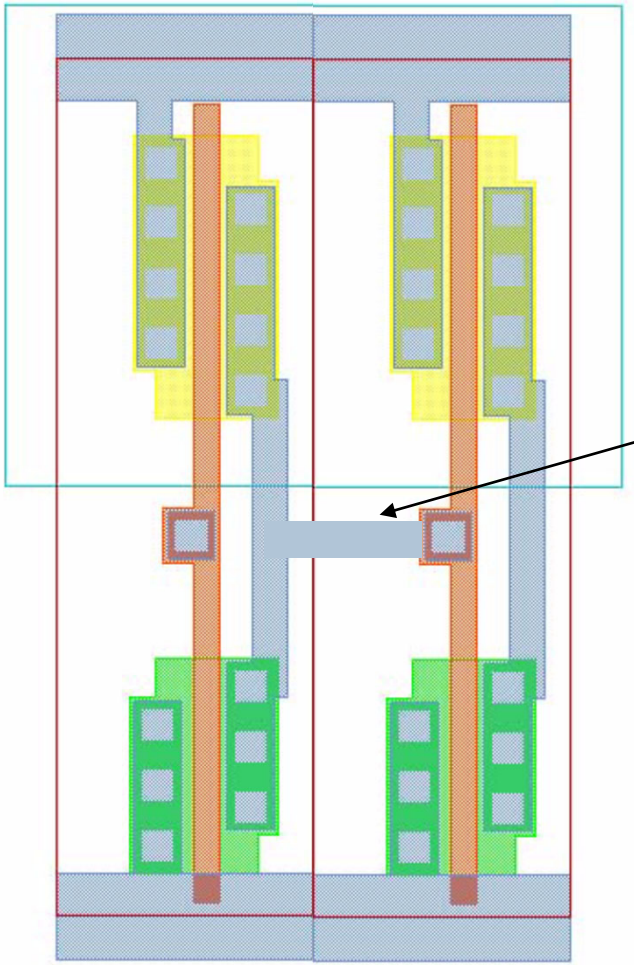


Two Inverters

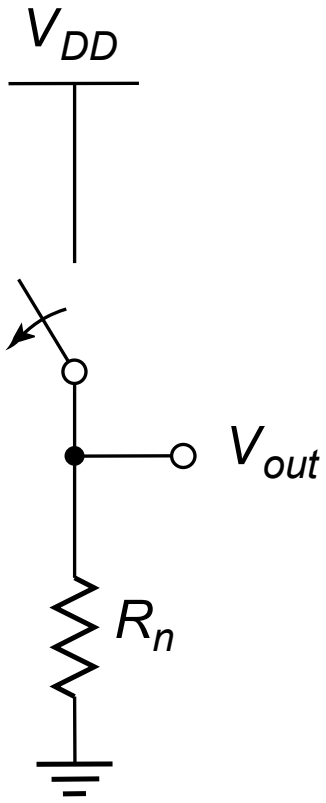
Share power and ground

Abut cells

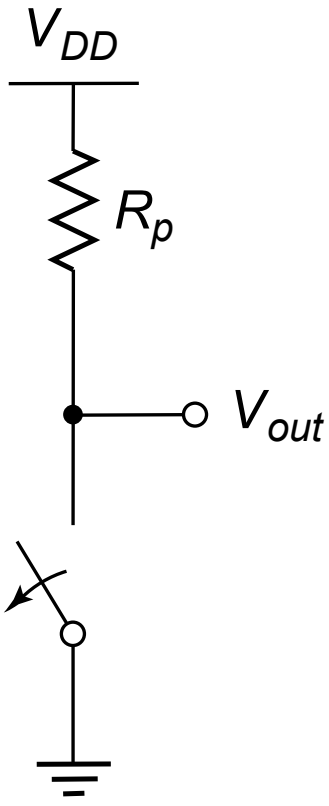




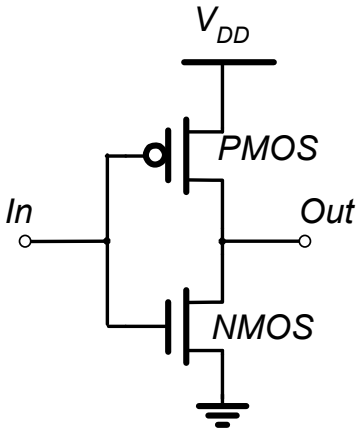
CMOS Inverter
First-Order DC Analysis



$V_{in} = V_{DD}$



$V_{in} = 0$

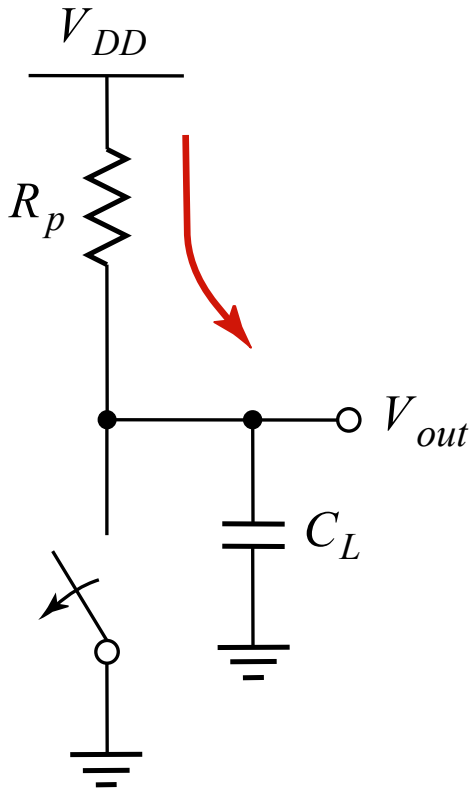


$V_{OL} = 0$
$V_{OH} = V_{DD}$
$V_M = f(R_n, R_p)$

基本稳态特性

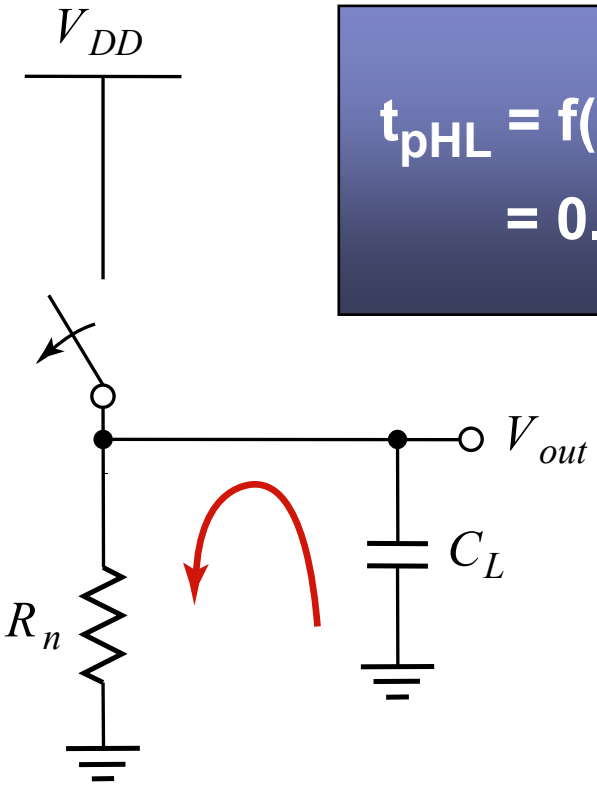
- 输出电压摆幅为V_{DD} ＝> 很高的噪声容限；
- 无比逻辑：V_{out}与晶体管的相对尺寸无关，所以可以采用最小尺寸；（有比逻辑：V_{out}由组成逻辑的晶体管的相对尺寸来决定。）
- 稳态时输出和V_{DD}或V_{SS}之间总存在一条具有电阻的通路＝>低输出电阻＝ >对噪声和干扰不敏感；
- 输入是晶体管的栅极＝>非常高的输入电阻＝>理论上，单个反相器可以驱动无穷多个门（无穷大的扇出），但大扇出会增加传播延时；
- 在V_{DD}和V_{SS}之间没有直接通路＝>无静态功耗；

CMOS Inverter: Transient Response



$V_{in} = 0$

(a) Low-to-high



$V_{in} = V_{DD}$

(b) High-to-low

$t_{pHL} = f(R_{on}.C_L)$
$= 0.69\ R_{on}C_L$

基本的动态特性

- 负载电容是输出节点所有电容之和；
- 传输时间由通过电阻对电容的充放电决定——高速门需要小的输出电容和导通电阻；
- 晶体管的尺寸（**W/L**）影响门的动态行为；

- 注意：**MOS**管的导通电阻不是一个常数值。

内容提要

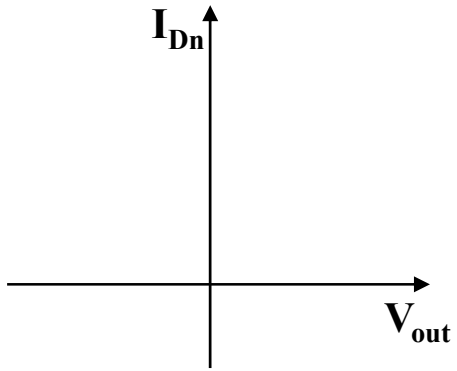
- 直观综述
- 电压传输特性（**VTC**）
- 可靠性：静态特性
- 性能：动态特性
- 功耗和能耗—延时积
- 按比例缩小技术以及对反相器的影响

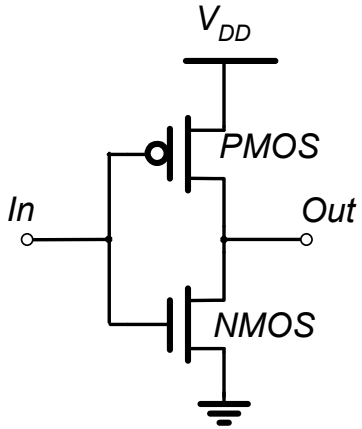
PMOS Load Lines

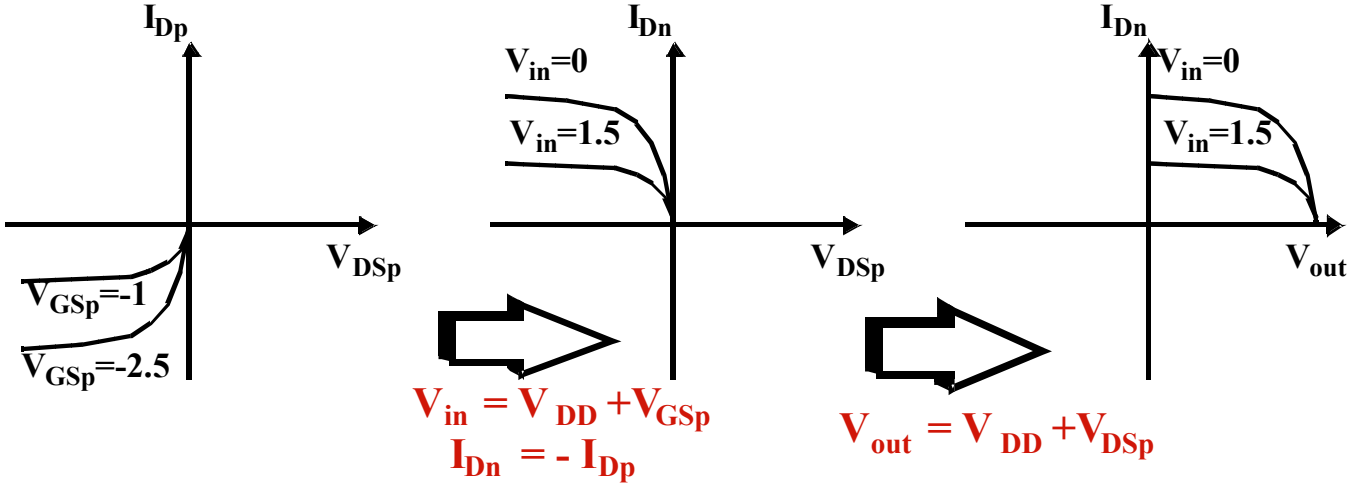
$$I_{Dp} = -I_{Dn}$$

$$V_{GSn} = V_{in}; \quad V_{GSp} = V_{in} - V_{DD}$$

$$V_{DSn} = V_{out}; \quad V_{DSp} = V_{out} - V_{DD}$$

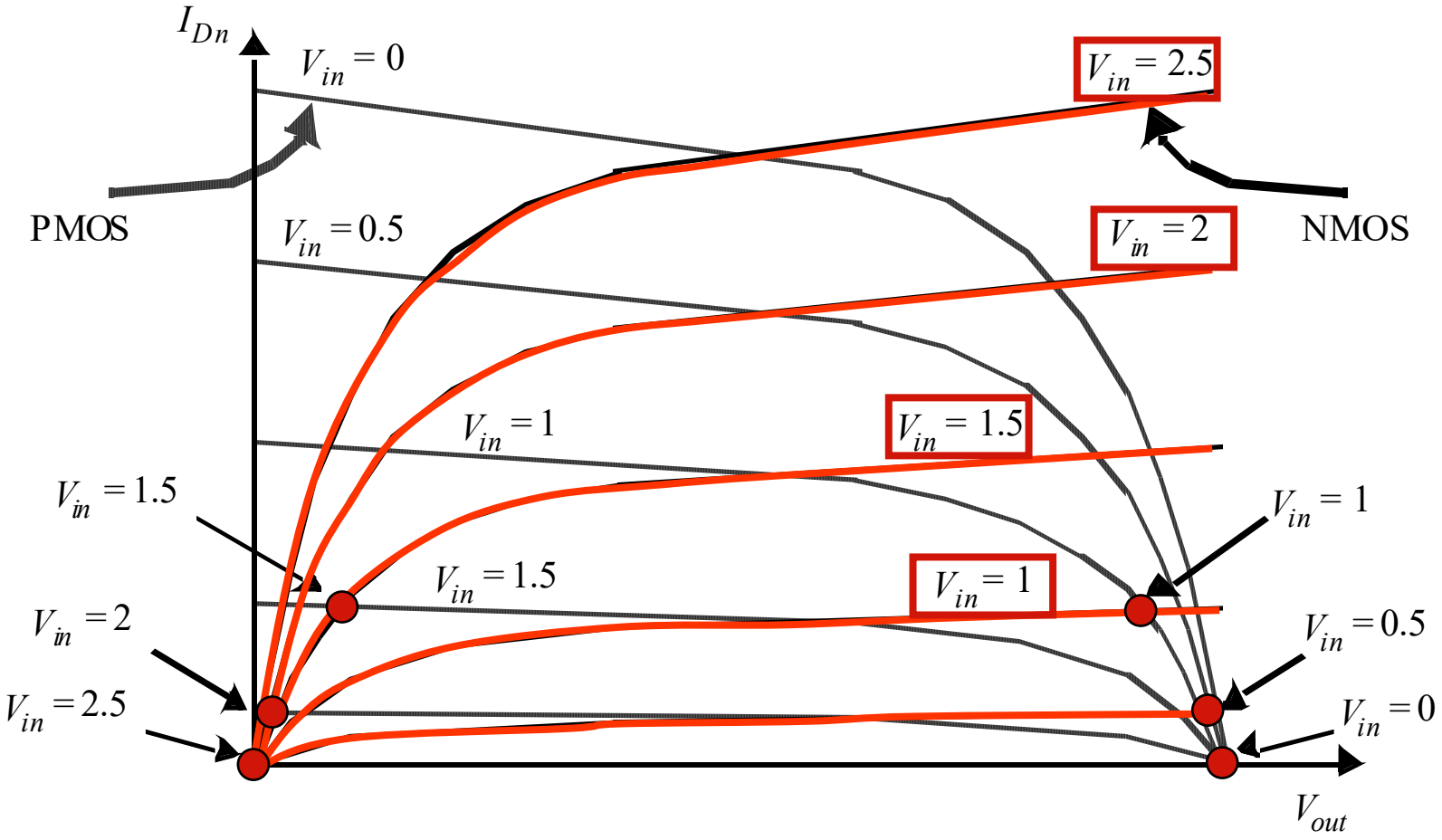






将PMOS I-V特性转换至公共坐标系（V_{DD}=2.5V）

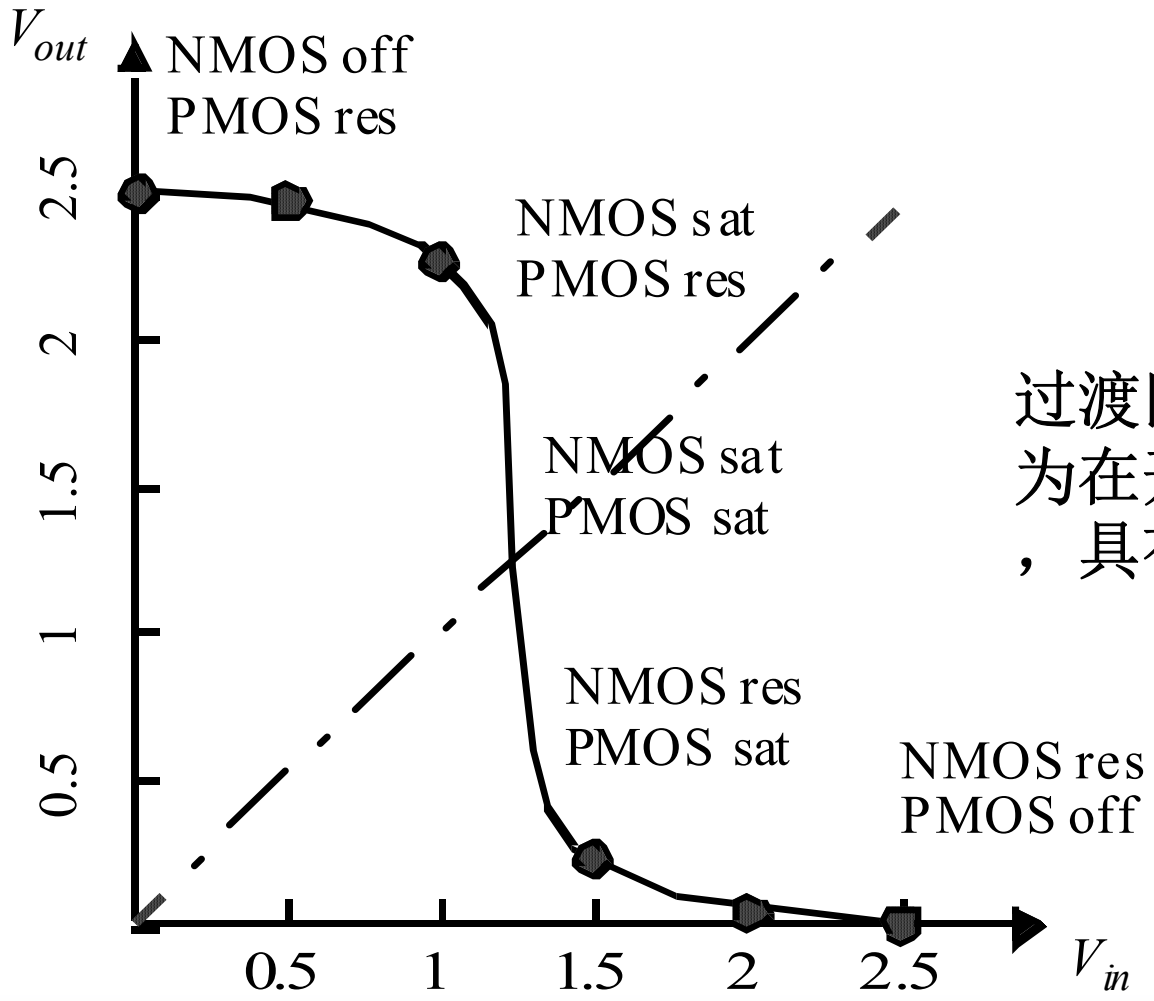
CMOS Inverter Load Characteristics



静态CMOS反相器中NMOS和PMOS管的负载曲线（VDD=2.5V）

圆点代表各种输入电压下的DC工作点。

CMOS Inverter VTC



过渡区非常窄，因为在开关过渡期间，具有高增益。

内容提要

- 直观综述
- 电压传输特性（**VTC**）
- 可靠性：静态特性
- 性能：动态特性
- 功耗和能耗－延时积
- 按比例缩小技术以及对反相器的影响

5.3.1 开关阈值

$V_{in}=V_{out}$

$V_{in}=V_M$时，两个晶体管均处于饱和状态，所以可以使用

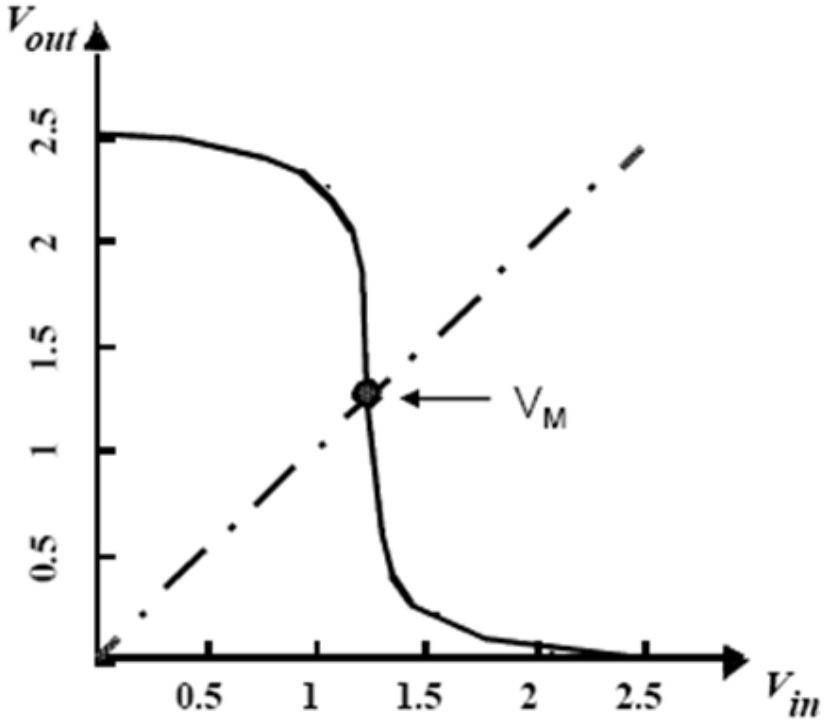
$I_n(V_{in}=V_M)=I_p(V_{in}=V_M)$ 来求解

晶体管饱和电流公式（3.38):

$$I_{DSAT}=v_{sat}C_{ox}W(V_{GS}-V_T-\frac{V_{DSAT}}{2})$$

$$=\frac{u_n V_{DSAT}}{L}C_{ox}W(V_{GS}-V_T-\frac{V_{DSAT}}{2})$$

$$=kV_{DSAT}(V_{GS}-V_T-\frac{V_{DSAT}}{2})$$



开关阈值的计算

$$I_n(V_{GS}=V_M)+I_p(V_{GS}=V_M-V_{DD})=0$$

$$k_n V_{DSATn}\left(V_M-V_{Tn}-\frac{V_{DSATn}}{2}\right)+k_p V_{DSATp}\left(V_M-V_{DD}-V_{Tp}-\frac{V_{DSATp}}{2}\right)=0$$

Solving for V_M yields

$$V_M=\frac{\left(V_{Tn}+\frac{V_{DSATn}}{2}\right)+r\left(V_{DD}+V_{Tp}+\frac{V_{DSATn}}{2}\right)}{1+r} \quad \text{with } r=\frac{k_p V_{DSATp}}{k_n V_{DSATn}}$$

$$V_M \approx \frac{rV_{DD}}{1+r}$$ 当V_{DD}值较大时

开关阈值取决于比值**r**，一般希望V_M处在$V_{DD}/2$附近，因此要求**r**接近于**1**。开关阈值等于所希望值时要求的**NMOS**和**PMOS**尺寸:

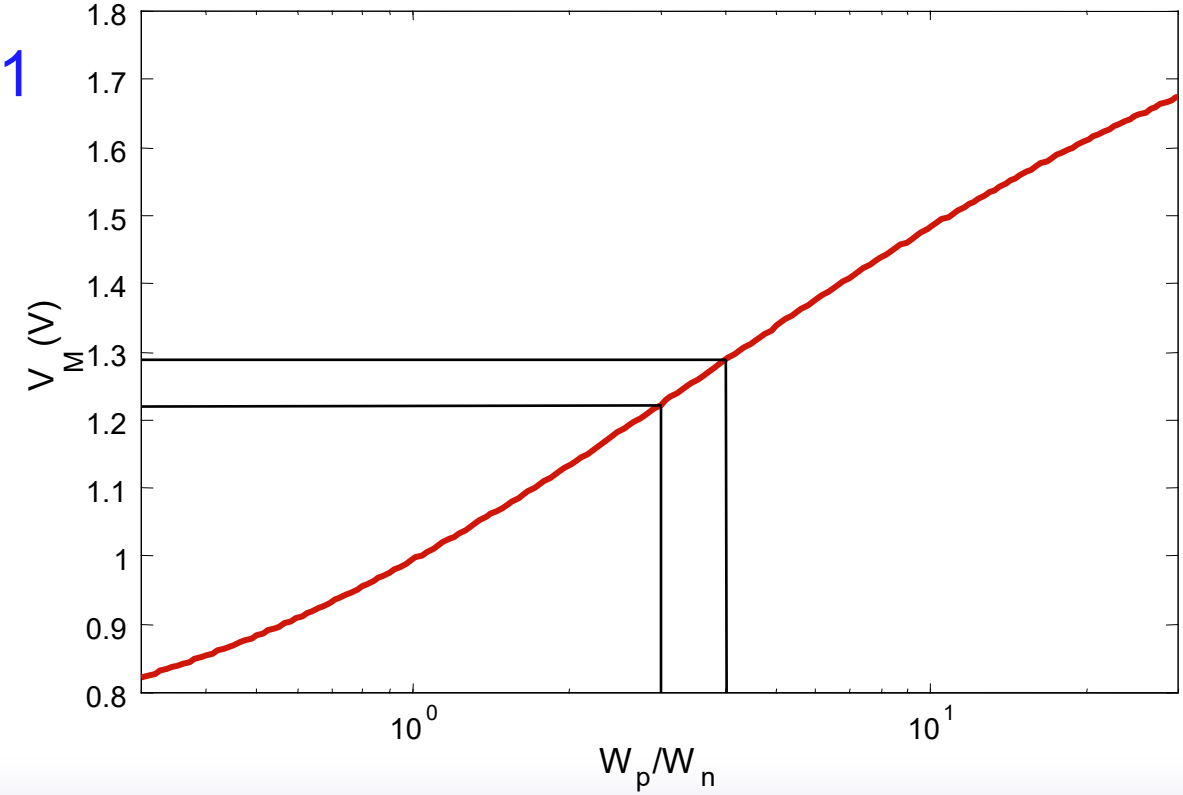
$$\frac{(W/L)_p}{(W/L)_n}=\frac{k_n' V_{DSATn}\left(V_M-V_{Tn}-V_{DSATn}/2\right)}{k_p' V_{DSATp}\left(V_{DD}-V_M+V_{Tp}-V_{DSATp}/2\right)}$$

0.25μm CMOS反相器的开关阈值

$$\frac{(W/L)_p}{(W/L)_n} = \frac{115\times10^{-6}\,A/V^2}{30\times10^{-6}\,A/V^2} \times \frac{0.63V}{1.0V} \times \frac{1.25-0.43-0.63/2}{1.25-0.4-1.0/2} = 3.5$$

V_M=1.25V　　仿真值：3.4

P135 例5.1



模拟反相器的开关阈值与PMOS对NMOS尺寸比的关系

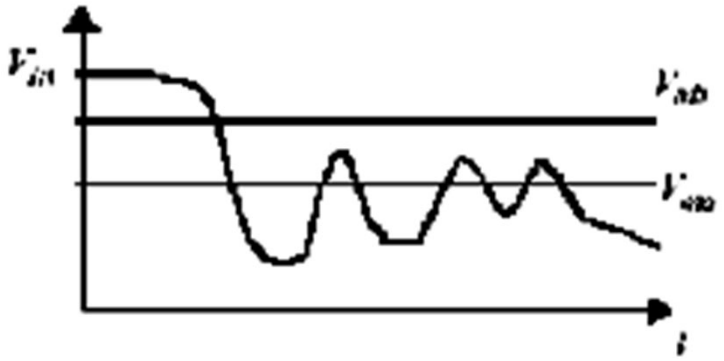
结论

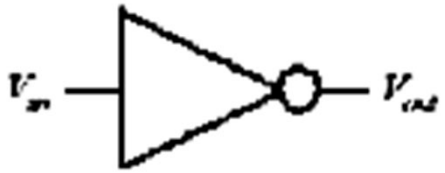
- V_M对于器件尺寸的比值变化不敏感
  - 例如，前面的例子中如果尺寸比为3、2.5和2，则V_M分别为1.22V、1.18V和1.13V

- 改变W_p对W_n比值的影响是使VTC的过渡区平移，增加PMOS或NMOS的宽度使V_M分别移向V_{DD}或GND.

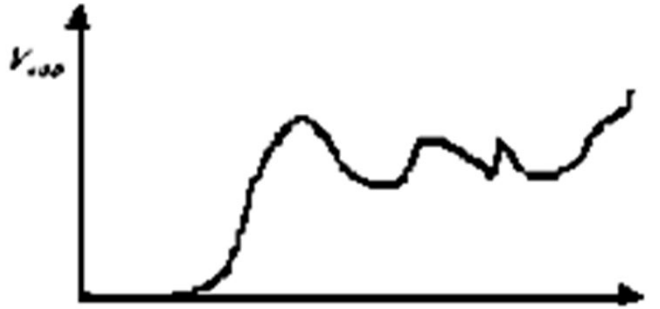
改变阈值的意义

- 通常要求(保证噪声容限) $V_M=V_{DD}/2$。
- V_M对于器件尺寸的比值变化不敏感。因此PMOS通常选择比$V_M=V_{DD}/2$所要求的尺寸小一些来减小面积。
- 在某些情况下并不总是要求$V_M=V_{DD}/2$。

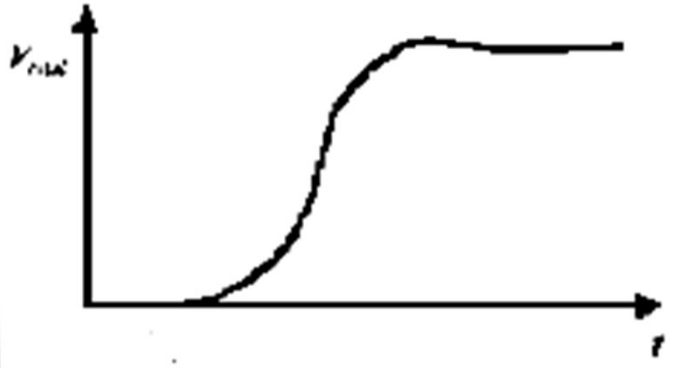




标准反相器的响应

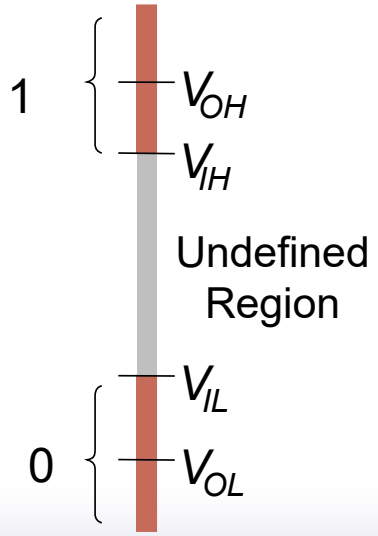


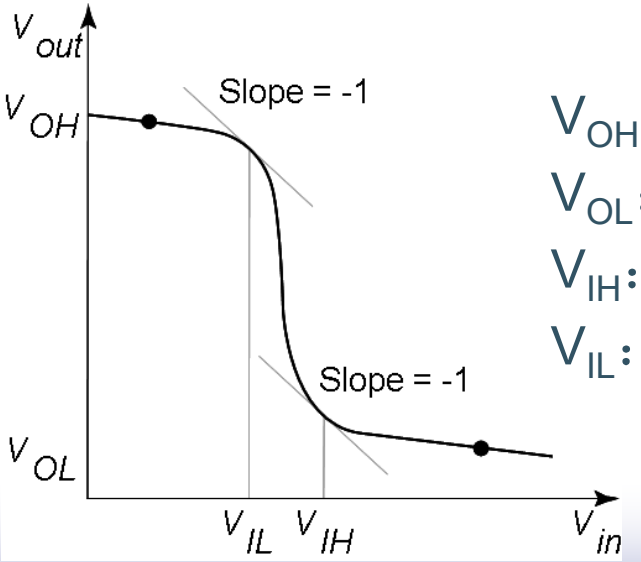
改变阈值后的反相器的响应



5.3.2 噪声容限

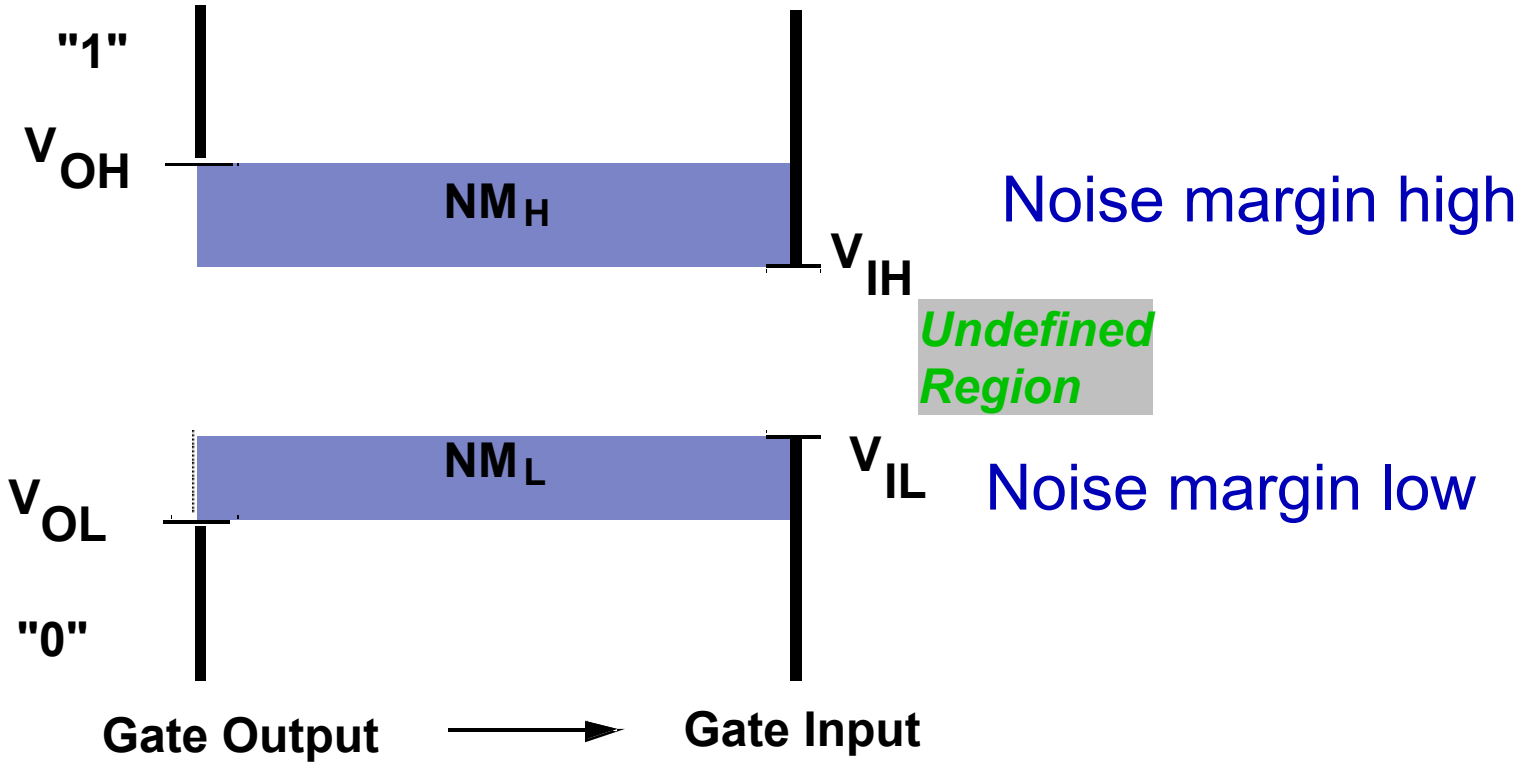
- 电路工作时，由于存在干扰信号，使输入电平偏离理想电平，影响输出电平；
- 用噪声容限反映电路的抗干扰能力。噪声容限反映电路能承受的实际输入电平与理想逻辑电平的偏离范围。





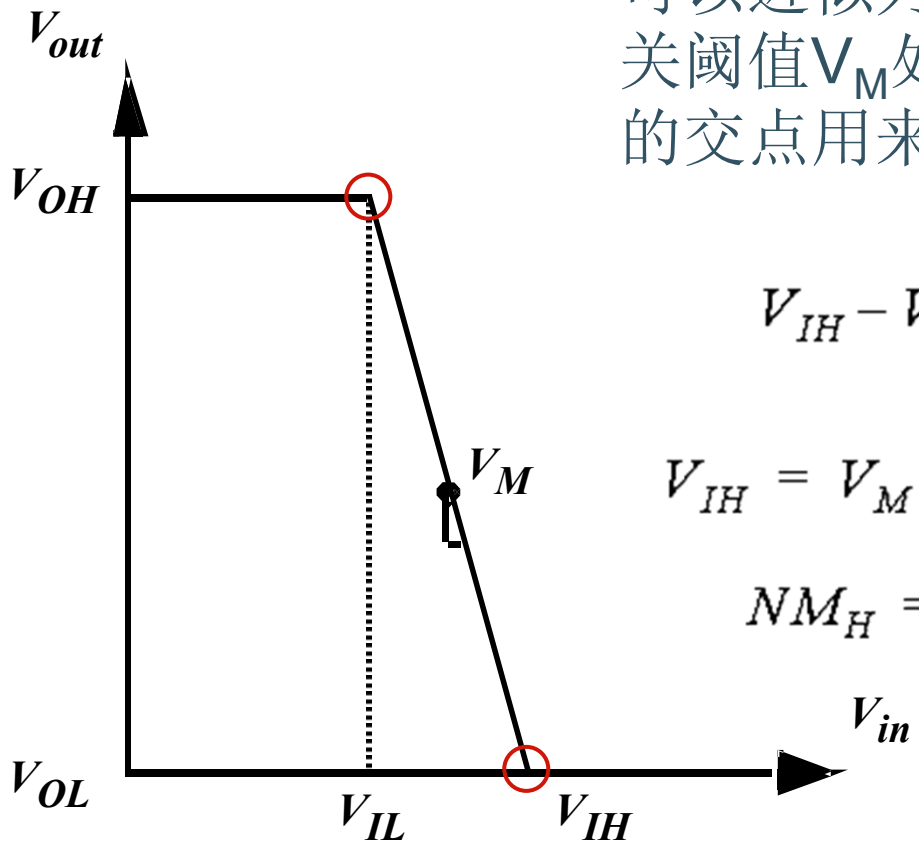
V_{OH}：额定高电平；
V_{OL}：额定低电平；
V_{IH}：可接受高电压；
V_{IL}：可接受低电压；

噪声容限的定义


"1"
V_{OH}
NM_H
V_{IH}
Noise margin high
Undefined Region
NM_L
V_{IL}
Noise margin low
V_{OL}
"0"
Gate Output
Gate Input

噪声容限V_{IH} and V_{IL}

对VTC采用逐段线性近似，过渡区可以近似为一段直线，其增益等于在开关阈值V_M处的增益g，它与V_{OH}和V_{OL}线的交点用来定义V_{IH}和V_{IL}点。



A simplified approach

$$V_{IH} - V_{IL} = -\frac{(V_{OH} - V_{OL})}{g} = \frac{-V_{DD}}{g}$$

$$V_{IH} = V_M - \frac{V_M}{g} \qquad V_{IL} = V_M + \frac{V_{DD} - V_M}{g}$$

$$NM_H = V_{DD} - V_{IH} \qquad NM_L = V_{IL}$$

Inverter Gain

在饱和区，增益与电流斜率关系很大，因此不能忽略沟道长度调制系数

$$k_n V_{DSATn}(V_{in}-V_{Tn}-\frac{V_{DSATn}}{2})(1+\lambda_n V_{out})$$

$$+k_p V_{DSATp}(V_{in}-V_{DD}-V_{Tp}-\frac{V_{DSATp}}{2})(1+\lambda_p V_{out}-\lambda_p V_{DD})=0$$

↓ 求导并求解

$$\frac{dV_{out}}{dV_{in}}=-\frac{k_n V_{DSATn}(1+\lambda_n V_{out})+k_p V_{DSATp}(1+\lambda_p V_{out}-\lambda_p V_{DD})}{\lambda_n k_n V_{DSATn}(V_{in}-V_{Tn}-\frac{V_{DSATn}}{2})+\lambda_p k_p V_{DSATp}(V_{in}-V_{DD}-V_{Tp}-\frac{V_{DSATp}}{2})}$$

↓ $V_{in}=V_M$，并忽略二次项

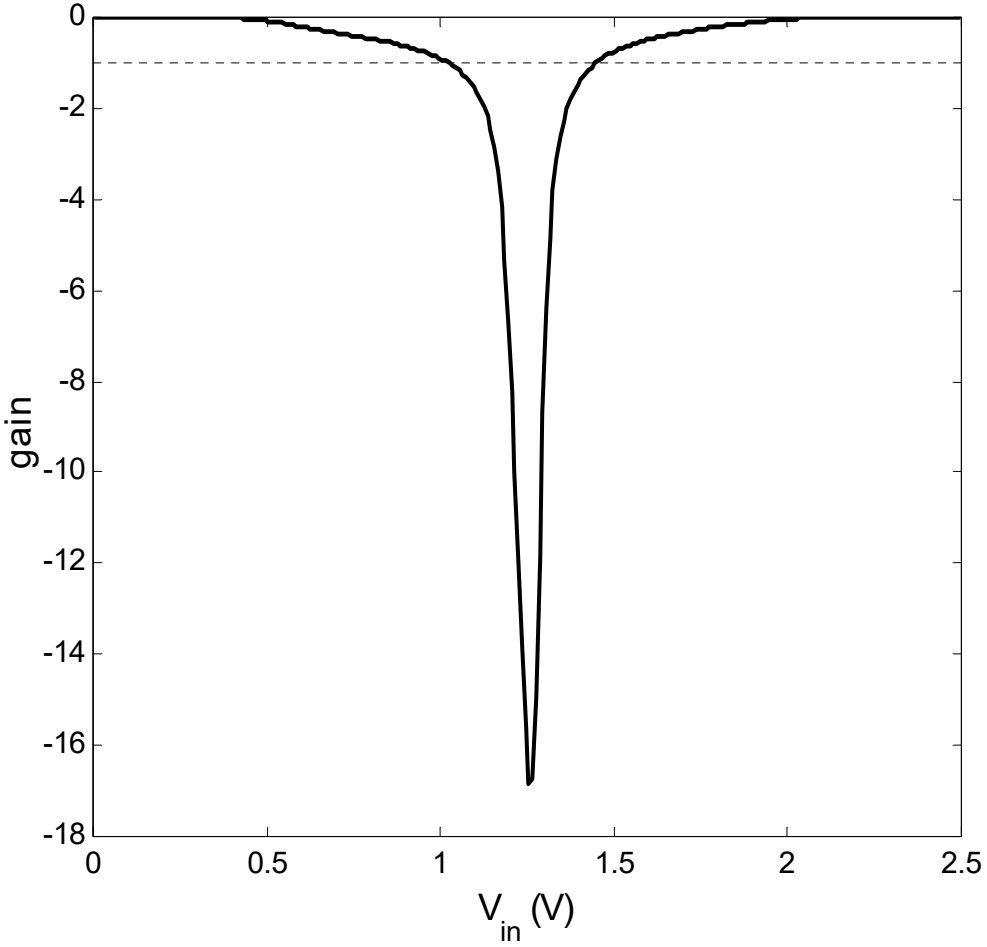
$$g=-\frac{1}{I_D(V_M)}\frac{k_n V_{DSATn}+k_p V_{DSATp}}{\lambda_n-\lambda_p}$$

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

关键参数是沟道长度调制参数**λ**;

V_{DD}和器件尺寸影响很小;

Inverter Gain

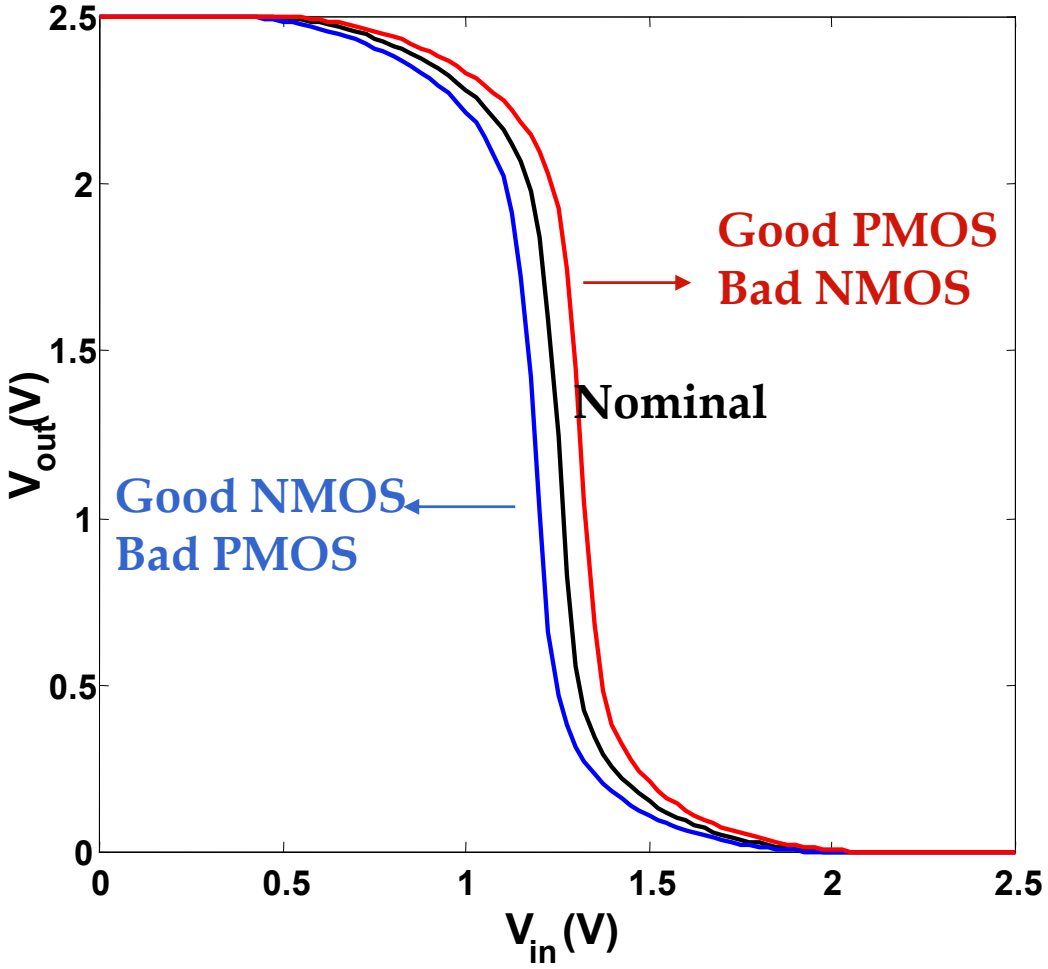


例5.2

$$g = -\frac{1}{I_D(V_M)}\frac{k_n V_{DSATn} + k_p V_{DSATp}}{\lambda_n - \lambda_p}$$

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

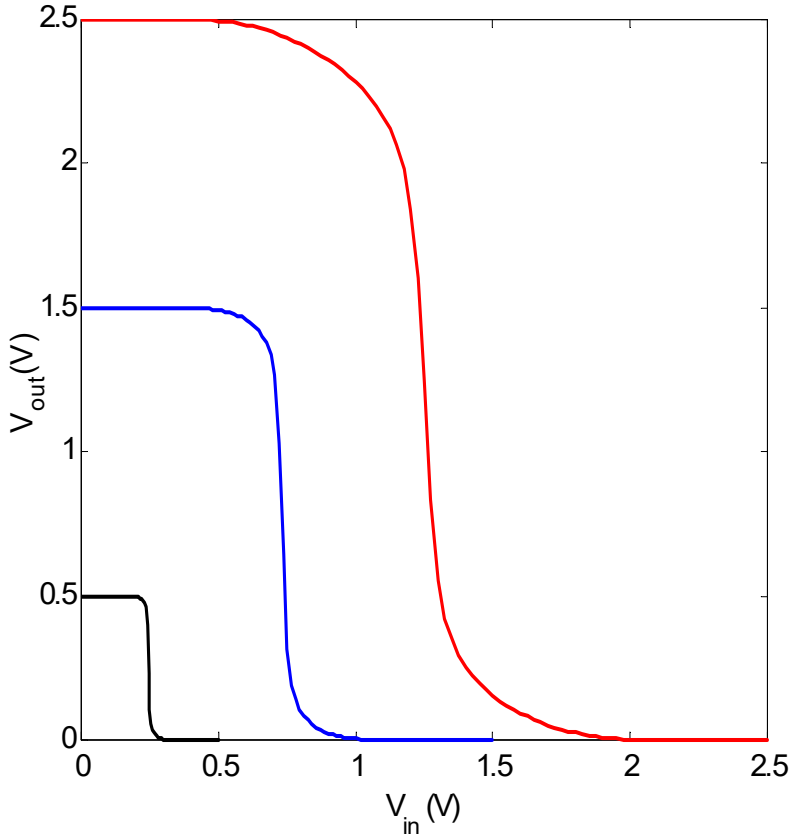
5.3.3 稳定性——器件参数变化

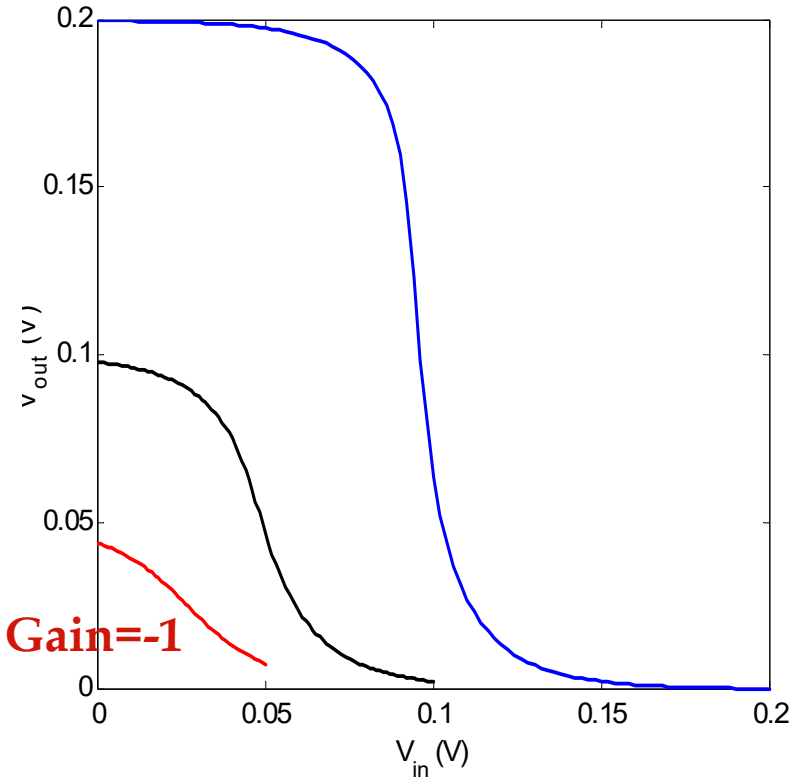


器件参数的变化使开关阈值平移，这一特性确保了门能在一个很宽范围的条件下工作，这也是静态**CMOS**门得以普遍使用的主要原因.

好的器件：具有较小的栅氧厚度、较小的长度、较大的宽度、较小的阈值。

稳定性—降低电源电压





反相器在过渡区的增益随电源电压降低而加大（公式**5.10**）

V_{DD}＝0.5V（只比晶体管的阈值高100mV时，过渡区宽度是电源电压的10％

V_{DD}＝2.5V时，加大到17％。降低V_{DD}可以改善反相器的直流特性。

降低V_{DD}的问题

- 不加区分的降低电源电压虽然对减少能耗有好处，但它会使门的延时加大；
- 一旦电源电压和阈值电压变得可以比拟，直流特性对器件参数的变化就变得越来越敏感；
- 降低电源电压意味着减小信号摆幅。虽然通常可以帮助减少系统的内部噪声（如由串扰引起的噪声），但它也使设计对并不减少的外部噪声源更加敏感。
- 电源电压至少等于热电势的两倍。

$$V_{DDmin} > 2...4\,\frac{kT}{q}$$

内容提要

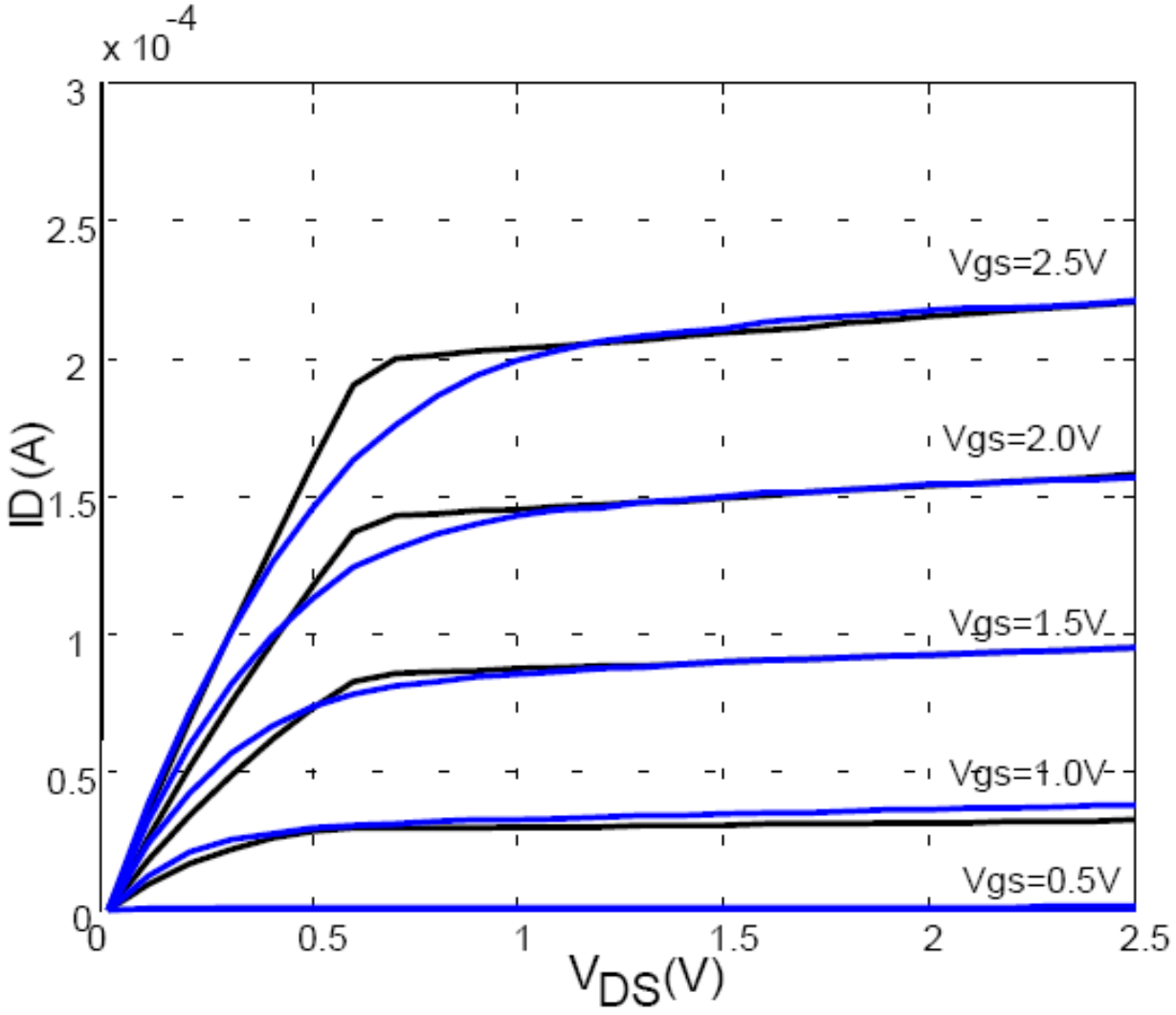
- 直观综述
- 电压传输特性（**VTC**）
- 可靠性：静态特性
- 性能：动态特性
- 功耗和能耗－延时积
- 按比例缩小技术以及对反相器的影响

5.4 动态特性

说明：

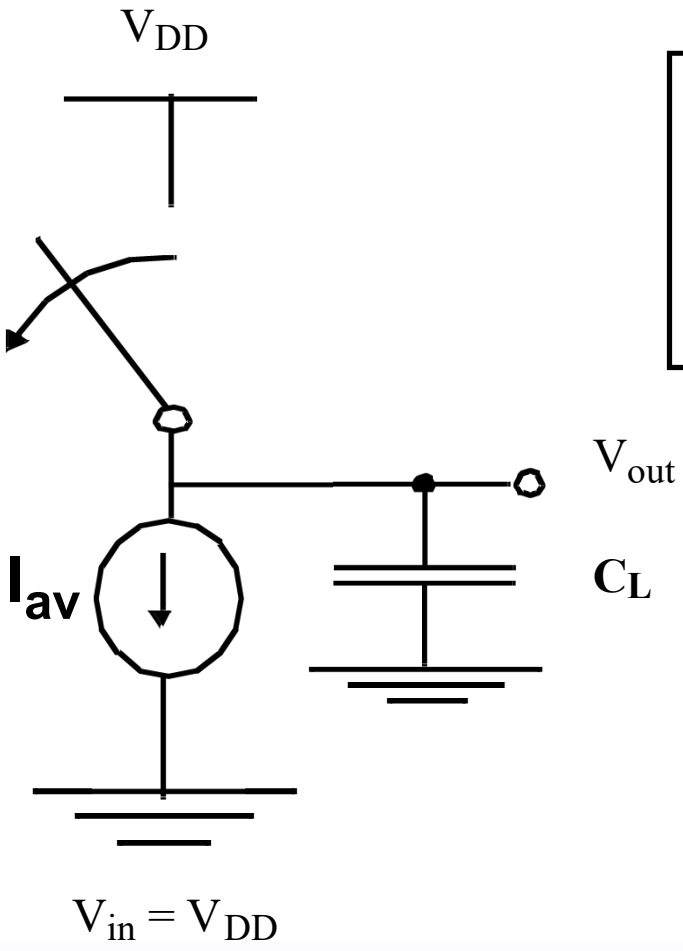
- 这里的计算结果只是一个近似值，大约有20～30%的偏差。更为精确的结果，可以使用SPICE仿真得到。
- 我们希望获得不同参数（尺寸，电阻，电容等）对设计性能影响的定性分析方法。

HSPICE和手工计算的对比



differences in the order of 10-20 %

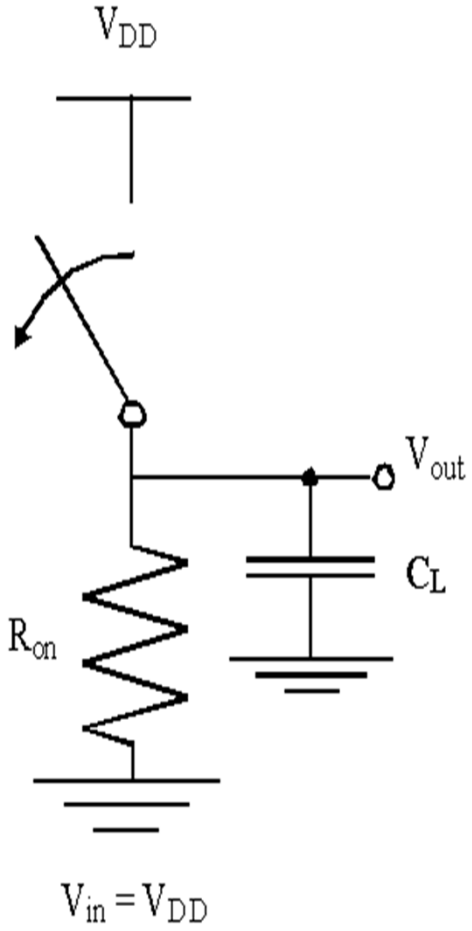
CMOS 反相器的传输延时——计算模型 1



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

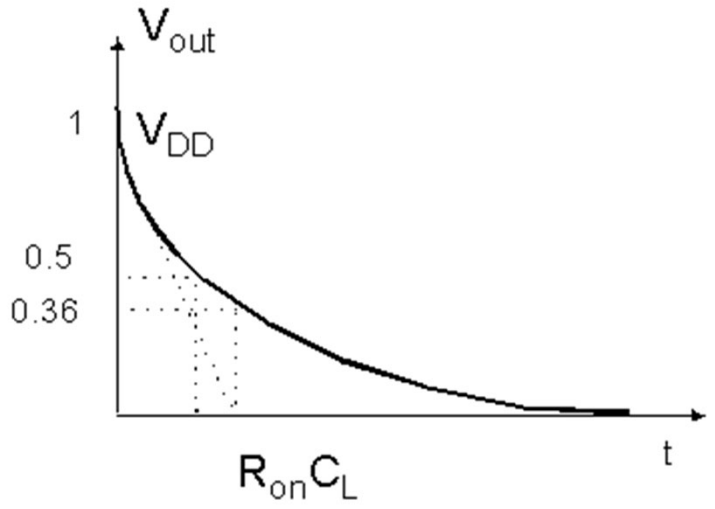
$$\approx \frac{C_L}{k_n \, V_{DD}}$$

CMOS 反相器的传输延时
——计算模型 2

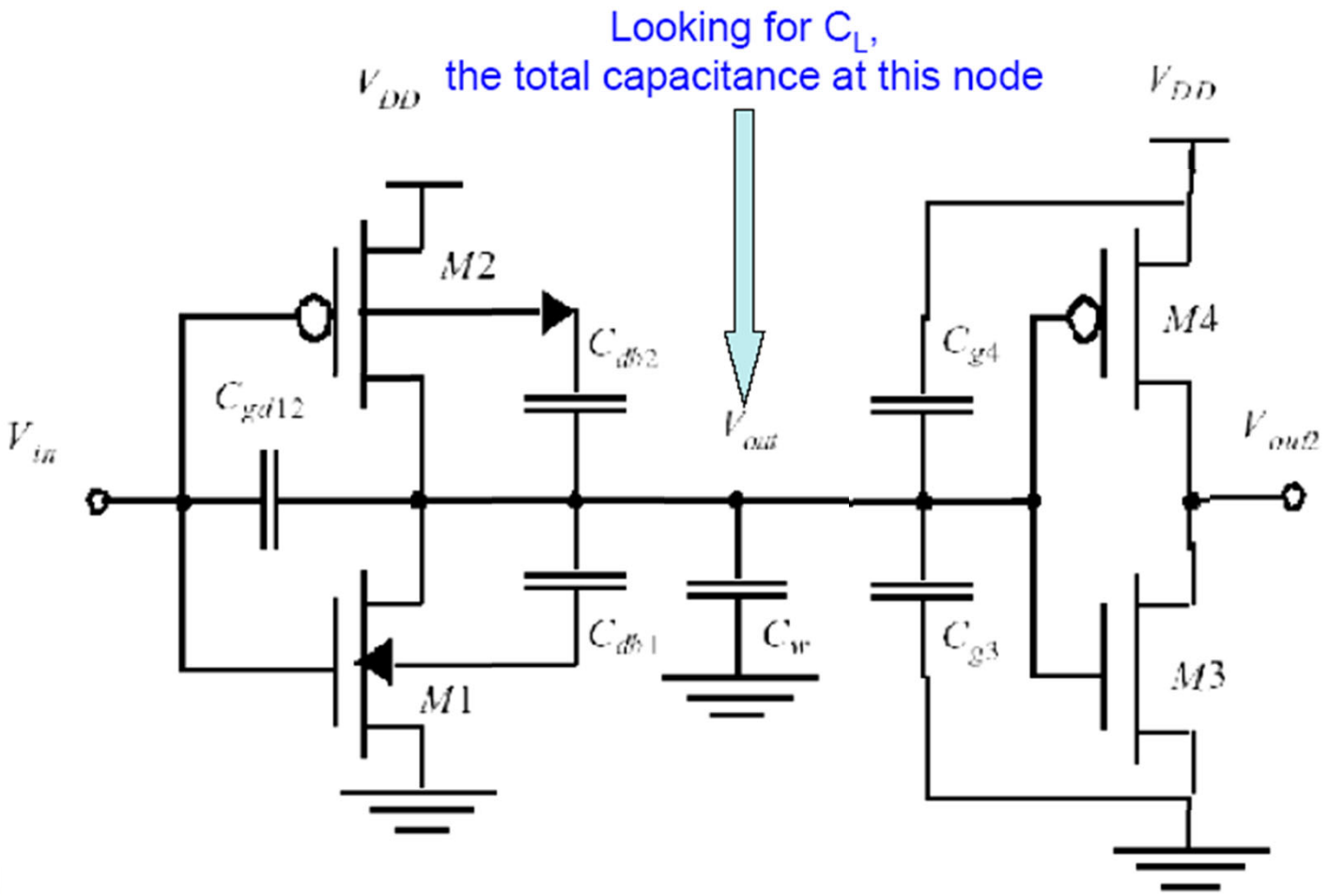


$$t_{pHL} = f(R_{on}.C_L)$$
$$= 0.69\, R_{on}C_L$$

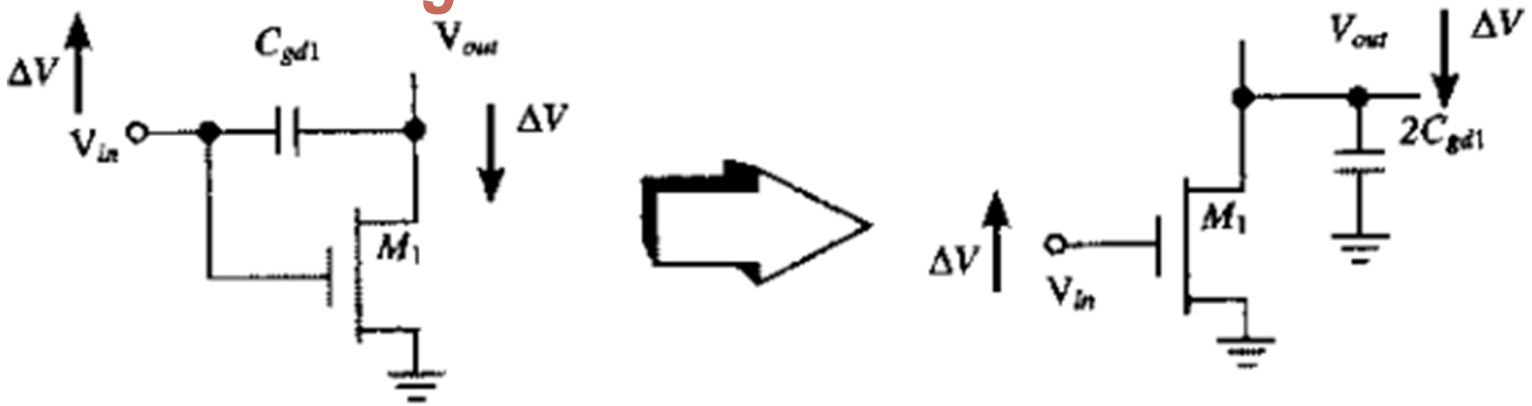
ln(0.5)

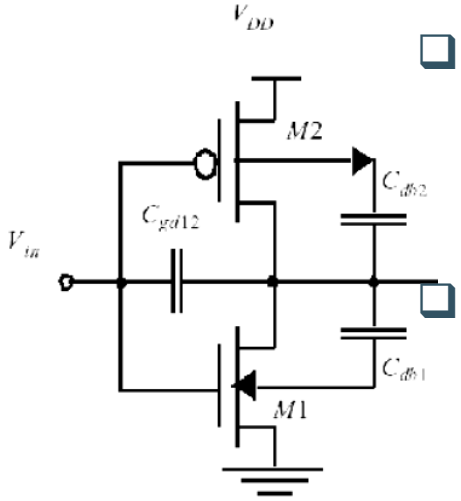


5.4.1 计算电容值


Looking for C_L,
the total capacitance at this node
V_{DD}
V_{DD}
M2
C_{db2}
C_{gd12}
V_{in}
V_{out}
C_{g4}
M4
V_{out2}
C_{db1}
C_w
C_{g3}
M1
M3

栅漏电容 C_{gd12}





- 在输出过渡的前半部分（至50%的点），晶体管M_1和M_2不是断开，就是饱和。C_{gd12}只包括M_1和M_2的覆盖电容.
- 在过渡期间，栅漏电容两端的电压向相反的方向变化，因此这一浮空电容上的电压变化是实际输出电压摆幅的两倍。为了在输出节点上出现同样的负载，接地电容必须是浮空电容的两倍：

$$C_{gd12} = 2C_{GD0}W$$

扩散电容 C_{db1} 和 C_{db2}

- 漏和体之间的电容来自反向偏置的pn结，这样的电容是高度非线性的。
- 使用线性电容替代。

P.59 3.10 3.11

$$C_{eq} = \frac{\Delta Q_j}{\Delta V_D} = \frac{Q_j(V_{high}) - Q_j(V_{low})}{V_{high} - V_{low}} = K_{eq} C_{j0}$$

$$K_{eq} = \frac{-\phi_0^m}{(V_{high} - V_{low})(1-m)} [(\phi_0 - V_{high})^{1-m} - (\phi_0 - V_{low})^{1-m}]$$

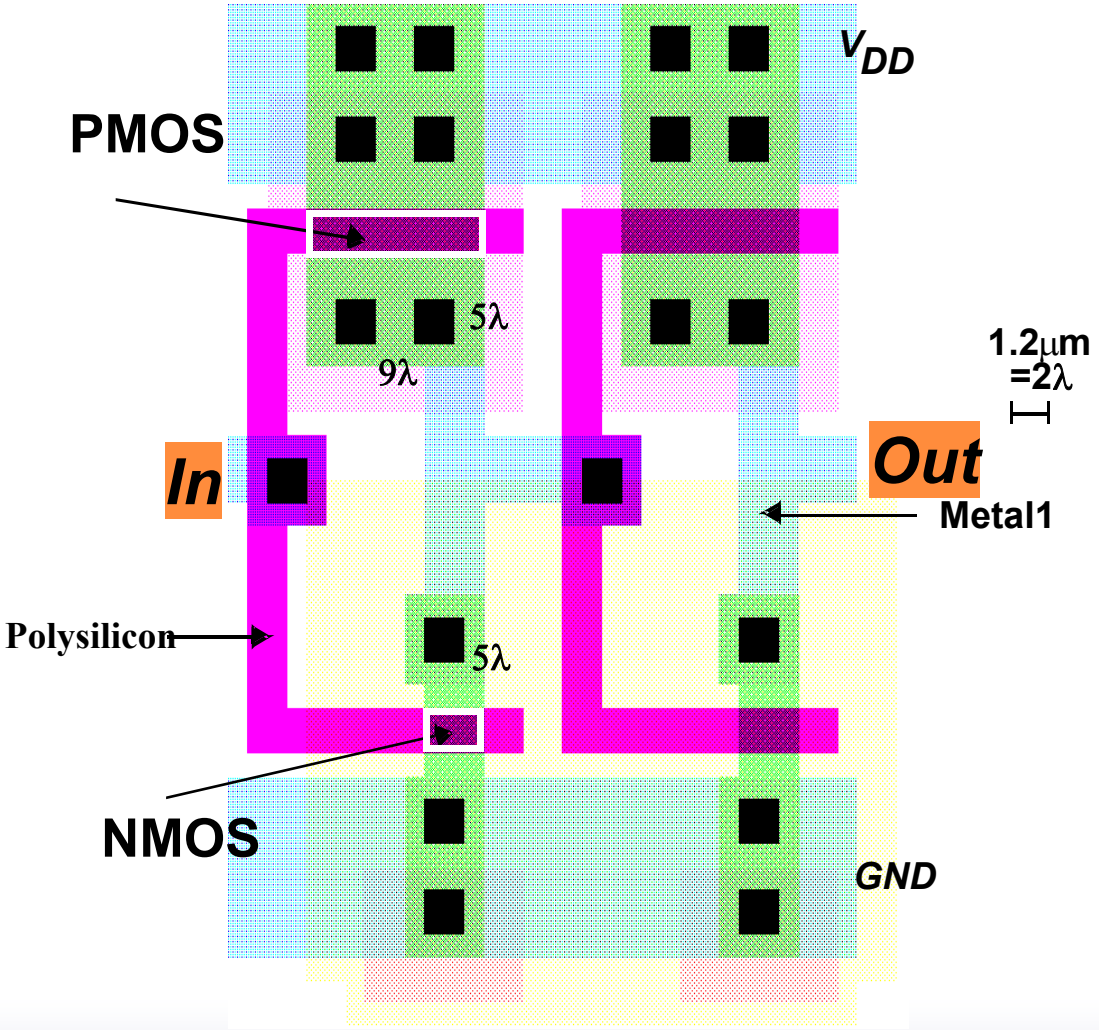
Φ_0是内建电势，m是结的梯度系数。

CMOS Inverters

P142 例5.3

P143 例5.4

扩散电容？？



连线电容 C_w

- 两个反相器之间的电容。由于两个反相器相邻（线很短），所以可以忽略不计；
- 如果连线很长，则不能忽略连线电容，需要从版图中提取具体数值；
- 以后的讲课内容会有专门一部分讲互连线及其对电路的影响；
- 现在工艺越来越先进，互连越来越重要。

扇出的栅电容C_{g3}和C_{g4}

$$C_{fan-out} = C_{gate}(NMOS) + C_{gate}(PMOS)$$
$$= \underbrace{(C_{GSOn} + C_{GDOn})} + W_n L_n C_{ox} + \underbrace{(C_{GSOp} + C_{GDOp})} + W_p L_p C_{ox}$$

覆盖电容

一个简化的表达式：

（1）忽略栅漏电容上的密勒效应。因为连接的门在达到50%点之前是不会翻转的。

（2）近似认为所连接的门的沟道电容在我们所关注的时间内保持不变。器件总的沟道电容为2/3WLC$_{ox}$（饱和），WLCox（线性或截止）。在过渡的前半段，其中一个负载器件一直处于线性模式，而另一个则从截止模式进入饱和模式。

电容—总结（1）

Capacitor	Expression	Remark
Cgd12 = 2(Cgdn+Cgdp)	$2(CGDO_n W_n + CGDO_p W_p)$	Gate drain capacitance of the first invertor with miller effect
Cgd34 = Cgdn+Cgdp	$(CGDO_n W_n + CGDO_p W_p)$	Gate drain capacitance of the second invertor (no miller effect)
Cdb1	$Keq_n AD_n CJ + Keqsw_n PDn CJSW$	Diffusion capacitance of nMOS drain junction of the first invertor
Cdb2	$Keq_p AD_p CJ + Keqsw_p PDn CJSW$	Diffusion capacitance of pMOS drain junction of the first invertor
Cg3	$CGSO_n W_n + Cox W_n L_n$ (1/2, 2/3, 1 ?)	Gate capacitance of nMOS of second invertor towards ground
Cg4	$CGSO_p W_p + Cox W_p L_p$	Gate capacitance of pMOS of second invertor towards Vdd
C_L	Sum of the above	

电容—总结（II）

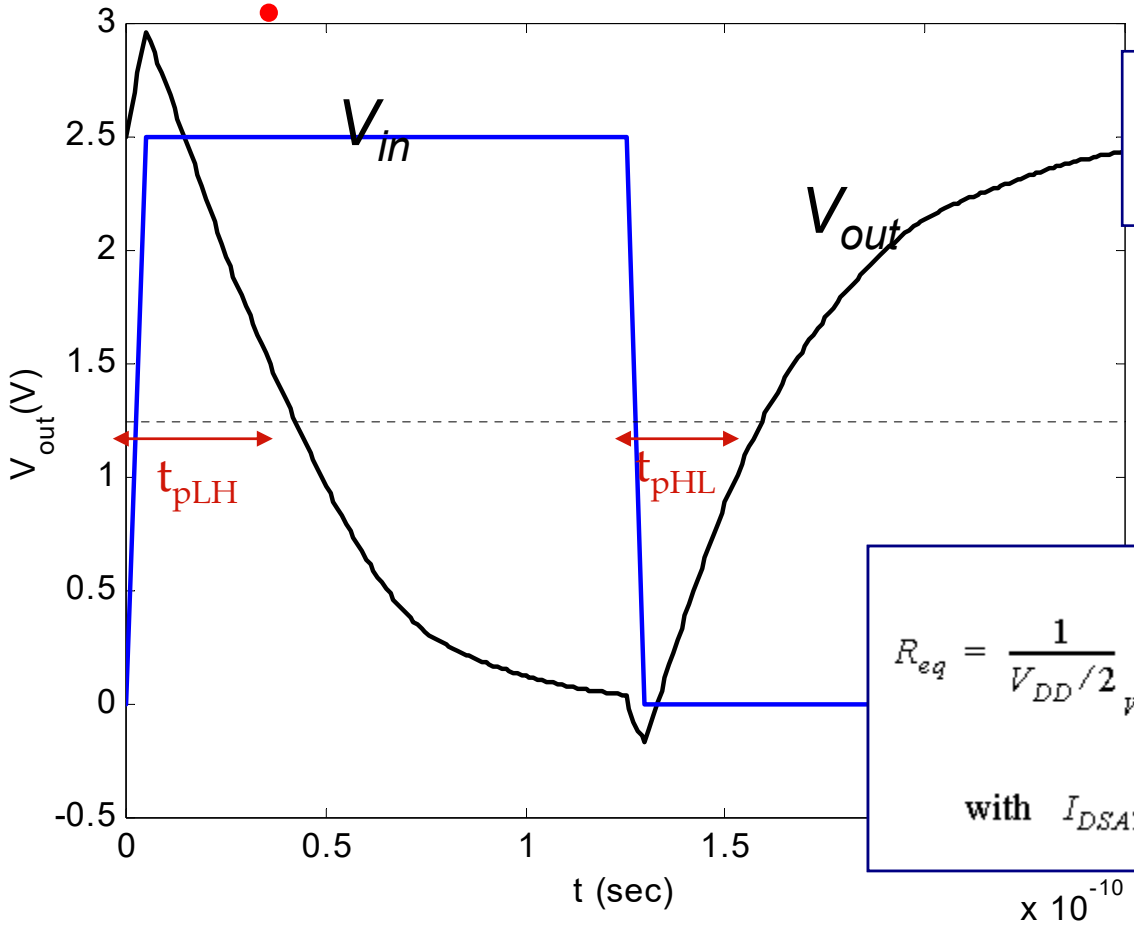
Table 5.2 Components of C_L (for high-to-low and low-to-high transitions).

Capacitor	Expression	Value (fF) (H→L)	Value (fF) (L→H)
C_{gd1}	$2\ CGD0_n\ W_n$	0.23	0.23
C_{gd2}	$2\ CGD0_p\ W_p$	0.61	0.61
C_{db1}	$K_{eqn}\ AD_n\ CJ + K_{eqswn}\ PD_n\ CJSW$	0.66	0.90
C_{db2}	$K_{eqp}\ AD_p\ CJ + K_{eqswp}\ PD_p\ CJSW)$	1.5	1.15
C_{g3}	$(CGD0_n+CGSO_n)\ W_n + C_{ox}\ W_n\ L_n$	0.76	0.76
C_{g4}	$(CGD0_p+CGSO_p)\ W_p + C_{ox}\ W_p\ L_p$	2.28	2.28
C_w	From Extraction	0.12	0.12
C_L	Σ	6.1	6.0

Transient Response

?

P.146 例5.5



$$t_p = (t_{pHL} + t_{pLH})/2 = 0.69\, C_L\, (R_{eqn} + R_{eqp})/2$$

$$R_{eq} = \frac{1}{V_{DD}/2} \int_{V_{DD}/2}^{V_{DD}} \frac{V}{I_{DSAT}(1+\lambda V)} dV \approx \frac{3}{4} \frac{V_{DD}}{I_{DSAT}} \left(1 - \frac{7}{9}\lambda V_{DD}\right)$$

$$\text{with} \quad I_{DSAT} = k' \frac{W}{L} \left((V_{DD} - V_T) V_{DSAT} - \frac{V_{DSAT}^2}{2} \right)$$

传播延时的计算公式

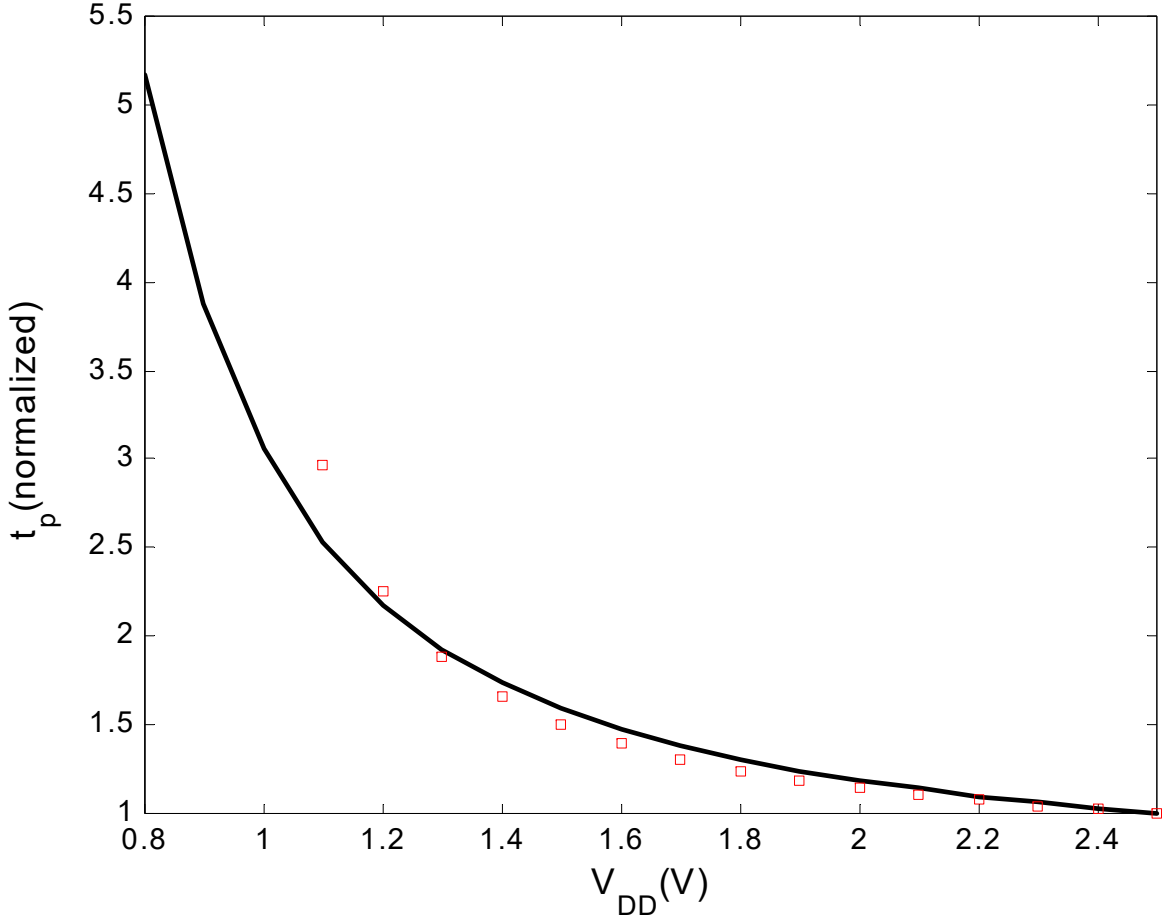
将式（5.18）和式（5.17）联合起来，忽略沟道调制效应，得到t_{pHL}的表达式：

$$t_{pHL} = 0.69 R_{eq} C_L = 0.69 \times \frac{3}{4} \times \frac{V_{DD}}{I_{DSATn}} \times C_L$$

$$= 0.52 \times \frac{C_L V_{DD}}{(W/L)_n k_n' V_{DSATn} (V_{DD} - V_{Tn} - V_{DSATn}/2)}$$

$$\approx 0.52 \times \frac{C_L}{(W/L)_n k_n' V_{DSATn}}$$

Delay as a function of V_{DD}



$$t_{pHL} = 0.69\frac{3}{4}\frac{C_L V_{DD}}{I_{DSATn}} = 0.52\frac{C_L V_{DD}}{(W/L)_n k'_n V_{DSATn}(V_{DD} - V_{Tn} - V_{DSATn}/2)}$$

减小门的传播延时的方法

- 减小C_L：精细的版图设计有助于减小扩散电容和互连线电容，优秀的设计实践要求漏扩散区的面积越小越好；
- 增加晶体管的**W/L**：这是设计者手中最有力和最有效的性能优化工具。
  - 但是增加晶体管的尺寸也增加扩散电容。
- 提高电源电压
  - 能量损耗。
  - 增加电源电压超过一定程度后改善就会非常有限，因而应当避免。
  - 从可靠性方面考虑，氧化层击穿和热载流子效应等问题迫使在深亚微米工艺中电源电压要规定严格上限。

5.4.3 设计综合考虑原则

NMOS与PMOS的比

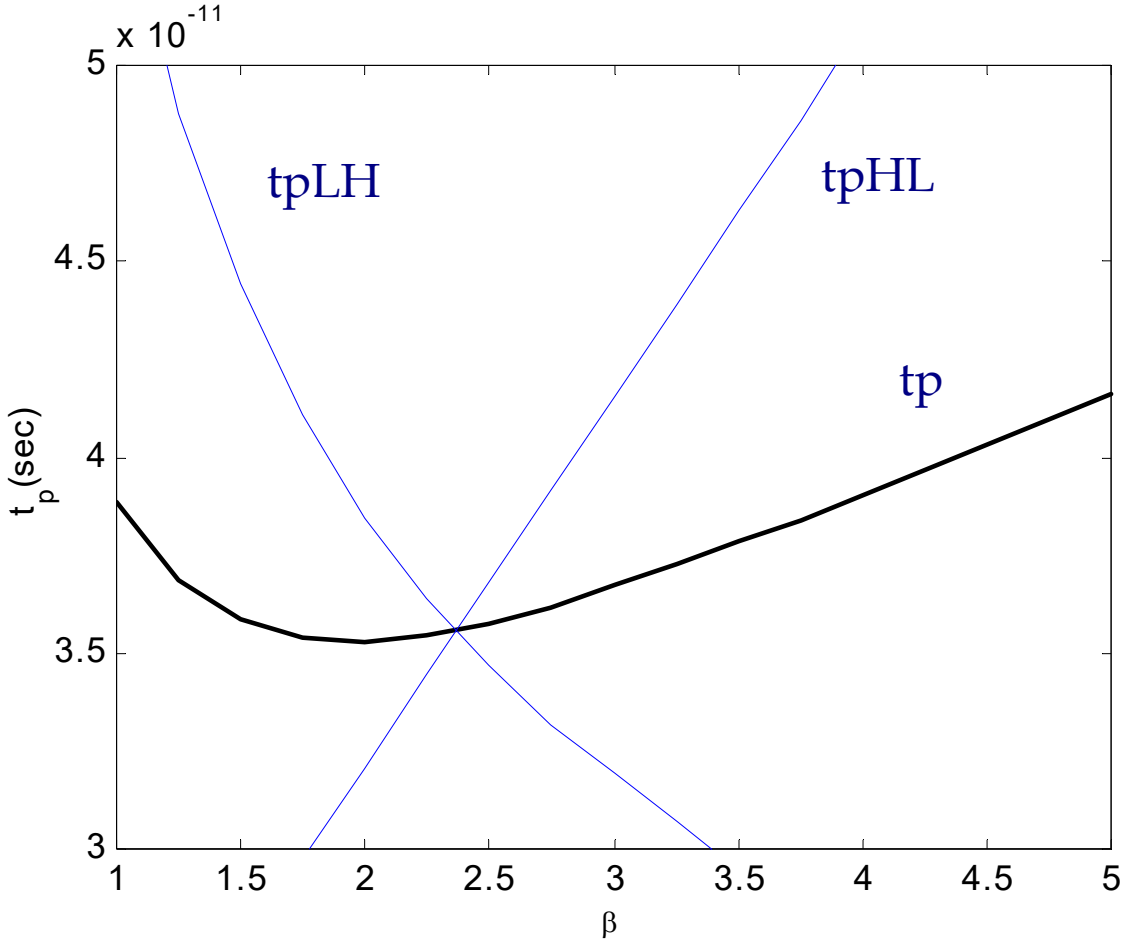
- PMOS较宽，使它的电阻与下拉的NMOS管匹配，这要求PMOS和NMOS的宽长比在3~3.5之间。采用这一方法得到对称的VTC，并且高至低与低至高的传播延时相等。**但是，总传播延时是否最小？？**
- 考虑到完全相同的两个反相器串联。$\beta=(W/L)_P/(W/L)_N$

$$C_L=(1+\beta)(C_{dn1}+C_{gn2})+C_w$$

$$t_p=\frac{0.69}{2}\left((1+\beta)(C_{dn1}+C_{gn2})+C_w\right)\left(R_{eqn}+\frac{R_{eqp}}{\beta}\right)$$

$$\beta_{opt}=\sqrt{r\left(1+\frac{C_w}{C_{dn1}+C_{gn2}}\right)} \qquad r=R_{eqp}/R_{eqn}$$

NMOS/PMOS ratio

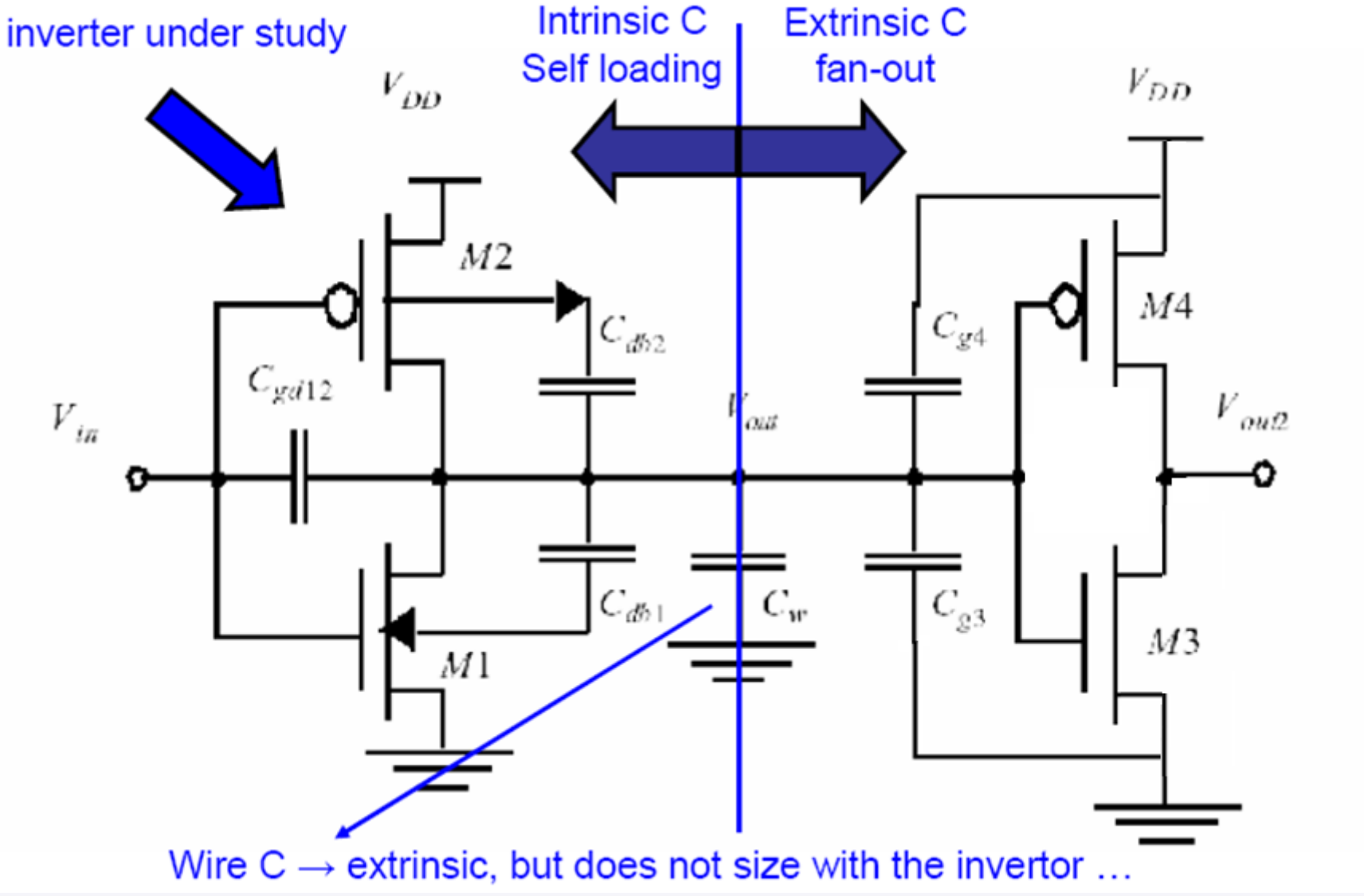


$$\beta=(W/L)_P/(W/L)_N$$

当$C_{dn1}+C_{gn2}>>C_w$时，

$$\beta_{opt}=\sqrt{r}$$

本征电容和外部电容


inverter under study
Intrinsic C
Self loading
Extrinsic C
fan-out
V_{DD}
V_{DD}
M2
C_{db2}
C_{gd12}
V_{in}
V_{out}
C_{g4}
M4
V_{out2}
C_{db1}
C_w
C_{g3}
M1
M3
Wire C → extrinsic, but does not size with the invertor …

本征延时与外部延时

- $C_L = C_{int} + C_{ext}$
- Invertor Delay:

$$t_p = 0.69R_{eq}(C_{int} + C_{ext}) = 0.69R_{eq}C_{int}(1 + \frac{C_{ext}}{C_{int}}) = t_{p0}(1 + \frac{C_{ext}}{C_{int}})$$

- Scaling of the invertor with a factor S:

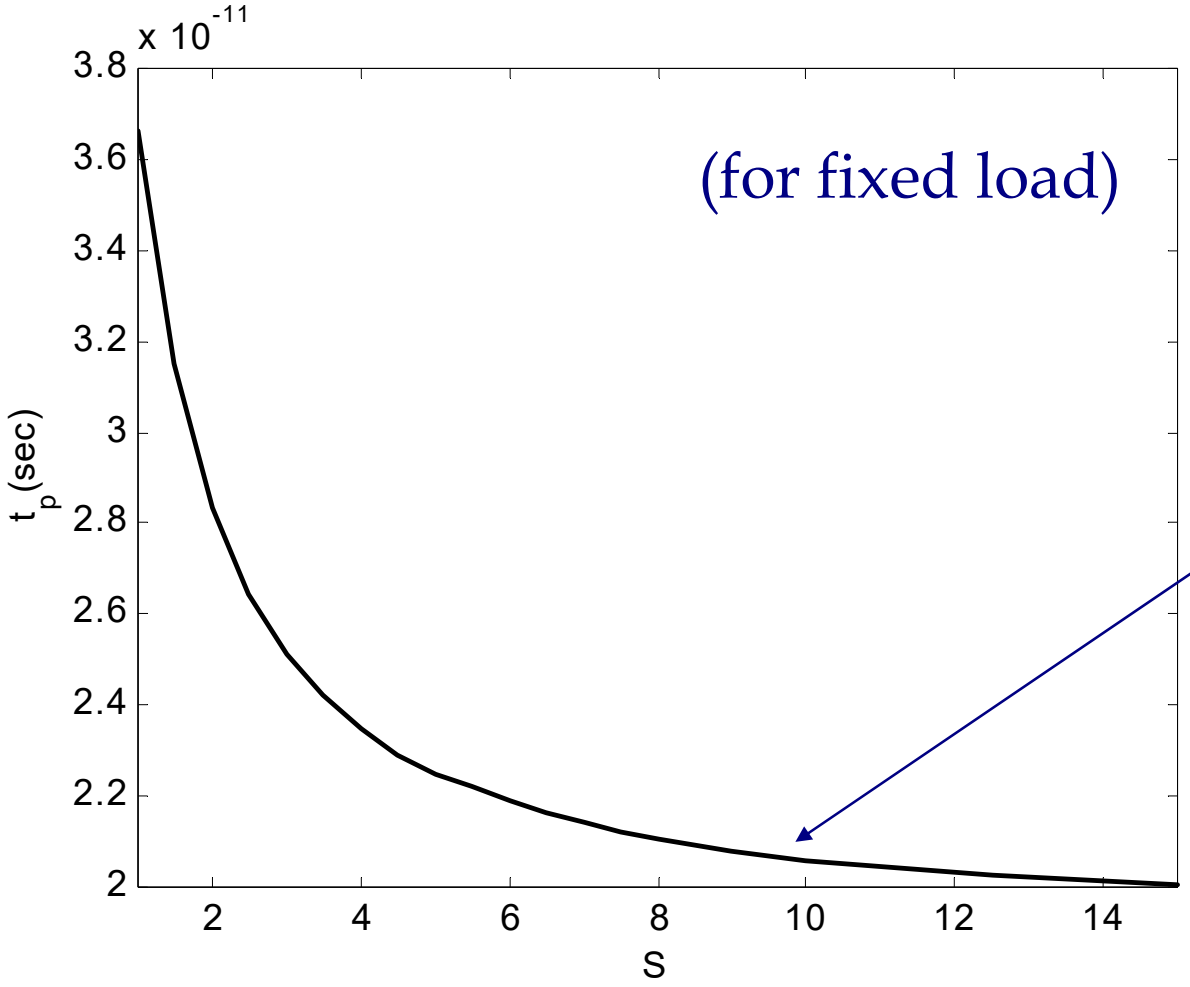
$$t_p = 0.69\frac{R_{ref}}{S}(SC_{iref})(1 + \frac{C_{ext}}{SC_{iref}}) = t_{p0}(1 + \frac{C_{ext}}{SC_{iref}})$$

1.反相器的本征延时t_{p0}与门的尺寸无关，而只取决于工艺以及版图。当不存在任何负载时，门的驱动强度的提高完全被随之增加的电容所抵消。

2.使S无穷大将达到最大的性能改善，因为消除了任何外部负载的影响，使延时减小到只有本征延时。然而足够大的尺寸系数S会显著增加硅面积而的到类似的结果。

Device Sizing

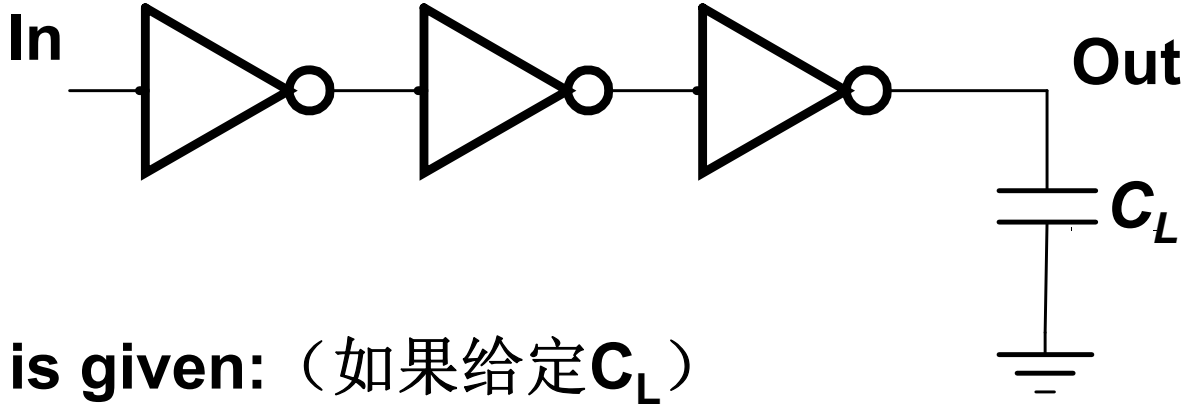
P150　例5.7



Self-loading effect:
Intrinsic capacitances dominate

真实电路中如何确定各级门尺寸?

反相器链



If C_L is given:（如果给定C_L）

- How many stages are needed to minimize the delay?
（需要多少级逻辑可以使延时最小？）

- How to size the inverters?
（反相器之间的大小关系怎样？）

May need some additional constraints.

反相器链的尺寸

输入栅电容与本征输出电容的关系：

$$C_{int} = \gamma C_g \qquad (\gamma \text{ is a tech. constant}, \approx 1)$$

外部负载电容即为下一级反相器的输入电容，并与尺寸成正比。

$$C_{ext} = fC_g$$ ***f***为等效扇出，

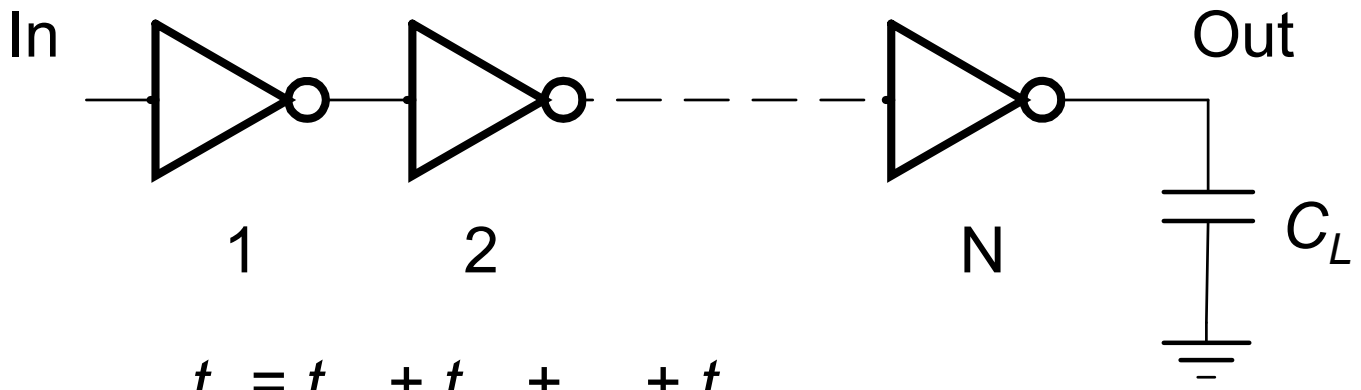
有效扇出

反相器的延时：

$$t_p = t_{p0}\left(1+\frac{C_{ext}}{C_{int}}\right) = t_{p0}\left(1+\frac{C_{ext}}{\gamma C_g}\right) = t_{p0}\left(1+\frac{f}{\gamma}\right)$$

反相器的延时只取决于外部负载电容与输入电容间的比值，即等效扇出***f***。

Apply to Inverter Chain



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

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

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

Optimal Tapering for Given N

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

Delay equation has N - 1 unknowns, $C_{gin,2} – C_{gin,N}$

N－1个未知数：$C_{g,2},\ldots,C_{g,N}$

Minimize the delay, find N - 1 partial derivatives

为了得到最小延时，通过求N－1次偏微分，并都等于0

Result: $C_{gin,j+1}/C_{gin,j} = C_{gin,j}/C_{gin,j-1}$

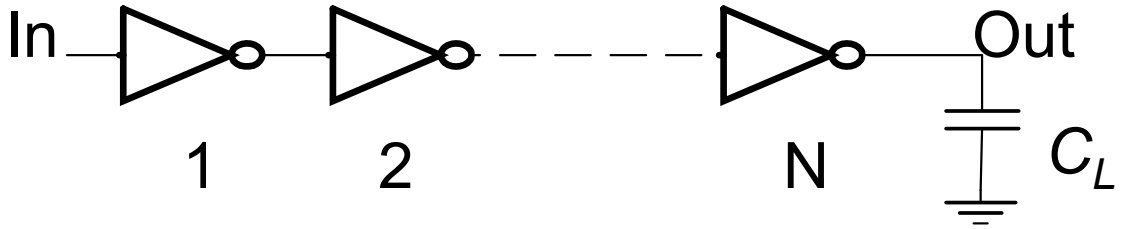
Size of each stage is the geometric mean of two neighbors

$$C_{gin,j} = \sqrt{C_{gin,j-1} C_{gin,j+1}}$$

每个反相器的最优尺寸是与它相临的两个反相器尺寸的几何平均数.

- each stage has the same effective fanout (C_{out}/C_{in})
- each stage has the same delay

优化的延时和门的级数



当C_{g1}和C_L已知时，则存在以下关系：

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

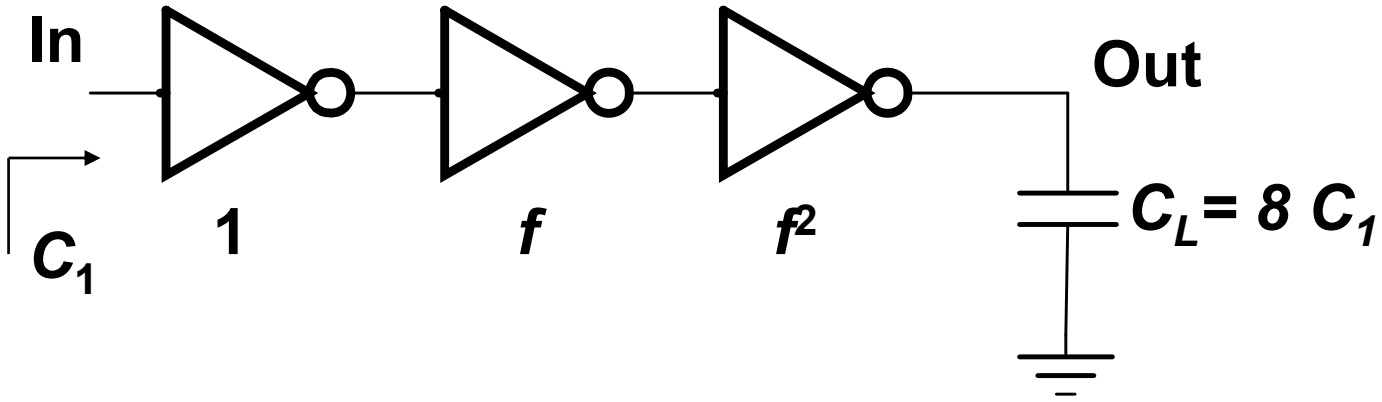
尺寸系数即等效扇出为:

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

反相器链的最小延时：

$$t_p = t_{p0}\sum_{i=1}^{N}\left(1+\frac{C_{gin,j+1}}{\gamma C_{gin,j}}\right) = Nt_{p0}\left(1+\sqrt[N]{F}/\gamma\right)$$

Example



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

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

Optimum Number of Stages

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

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

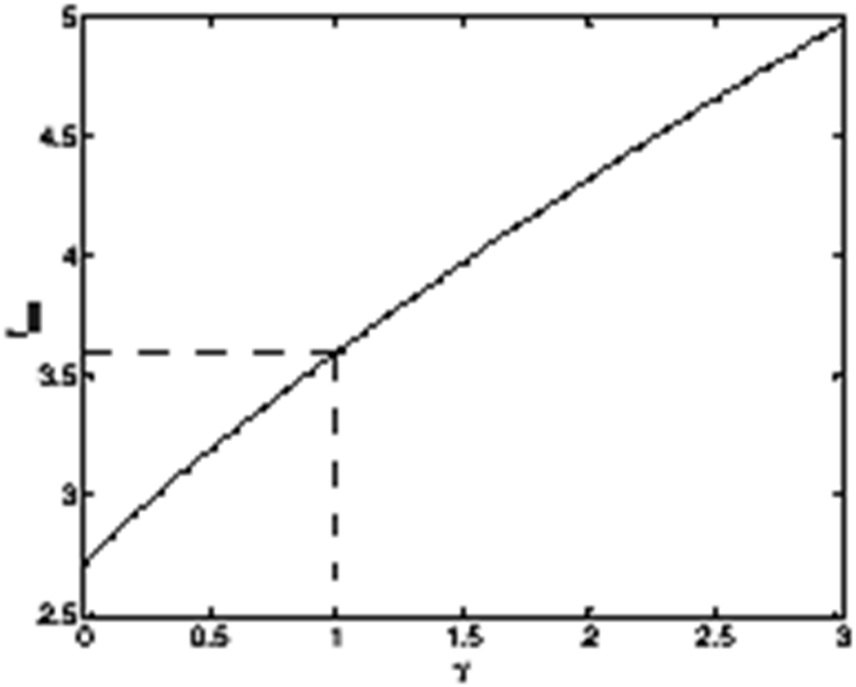
$$t_p = N t_{p0}\left(F^{1/N}/\gamma + 1\right) = \frac{t_{p0} \ln F}{\gamma}\left(\frac{f}{\ln f} + \frac{\gamma}{\ln f}\right)$$

$$\frac{\partial t_p}{\partial f} = \frac{t_{p0} \ln F}{\gamma} \cdot \frac{\ln f - 1 - \gamma/f}{\ln^2 f} = 0 \qquad f = \exp\left(1 + \gamma/f\right)$$

（1）当γ=0时，*f* = e，忽略自载得到的最优级数N=ln(F)，这一优化的缓冲器设计以一种指数形式逐级放大各级尺寸，并且成为指数锥形。

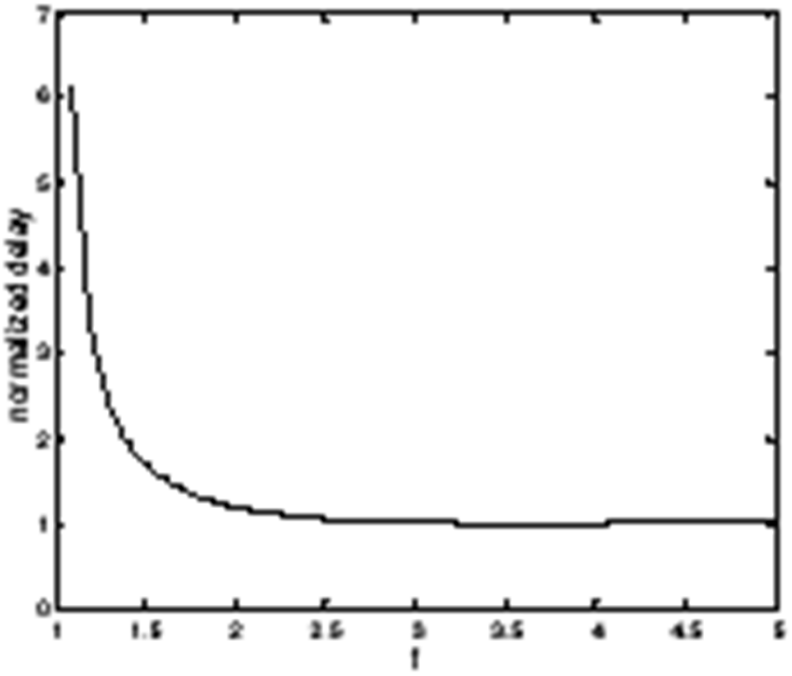
Optimum Effective Fanout f

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



f_{opt} = 3.6 for γ=1

Impact of Self-Loading on tp



With Self-Loading γ=1

（2）当γ=1（典型值）时，最优的锥形系数将接近于3.6。选择扇出值大于最优值并不会过多的影响延时，但能减少所要求的缓冲器级数和面积。一个通常的做法是选择最优的扇出为4。

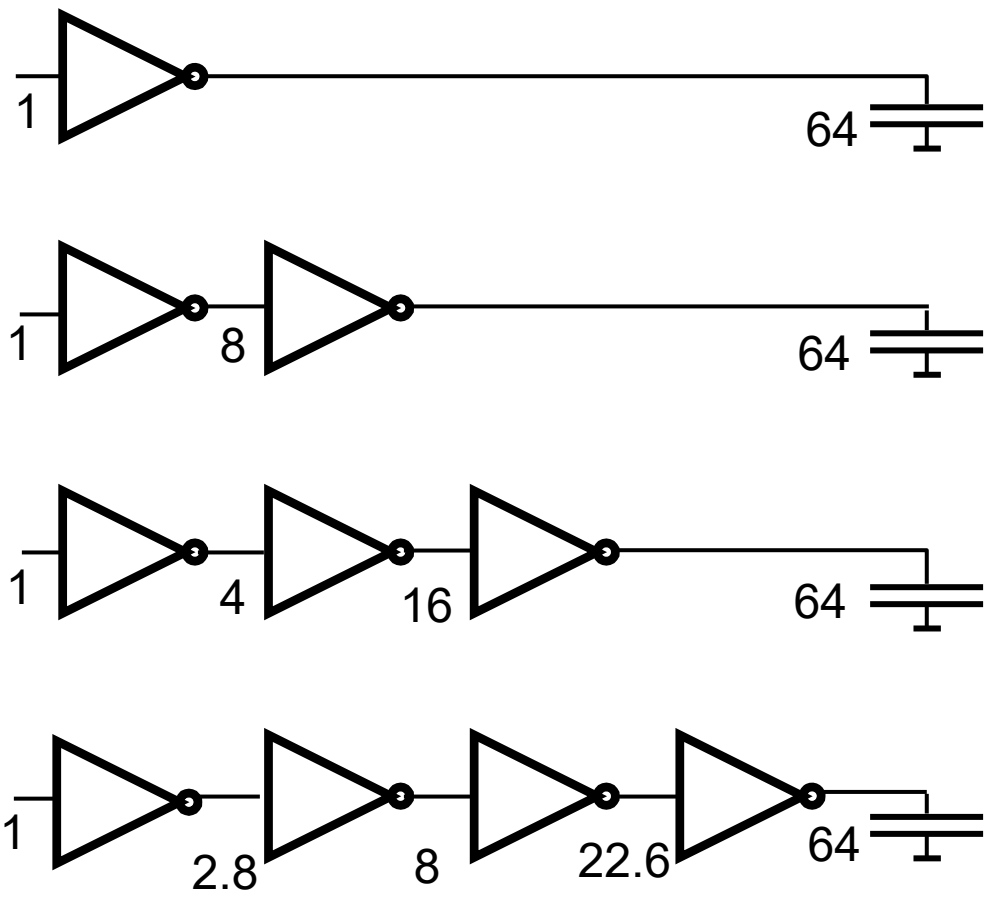
Normalized delay function of F

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

不同驱动器结构的t_{opt}/t_{p0}与F的关系

F	无缓冲器	两级反相器	反相器链
10	11	8.3	8.3
100	101	22	16.5
1000	1001	65	24.8
10000	10001	202	33.1

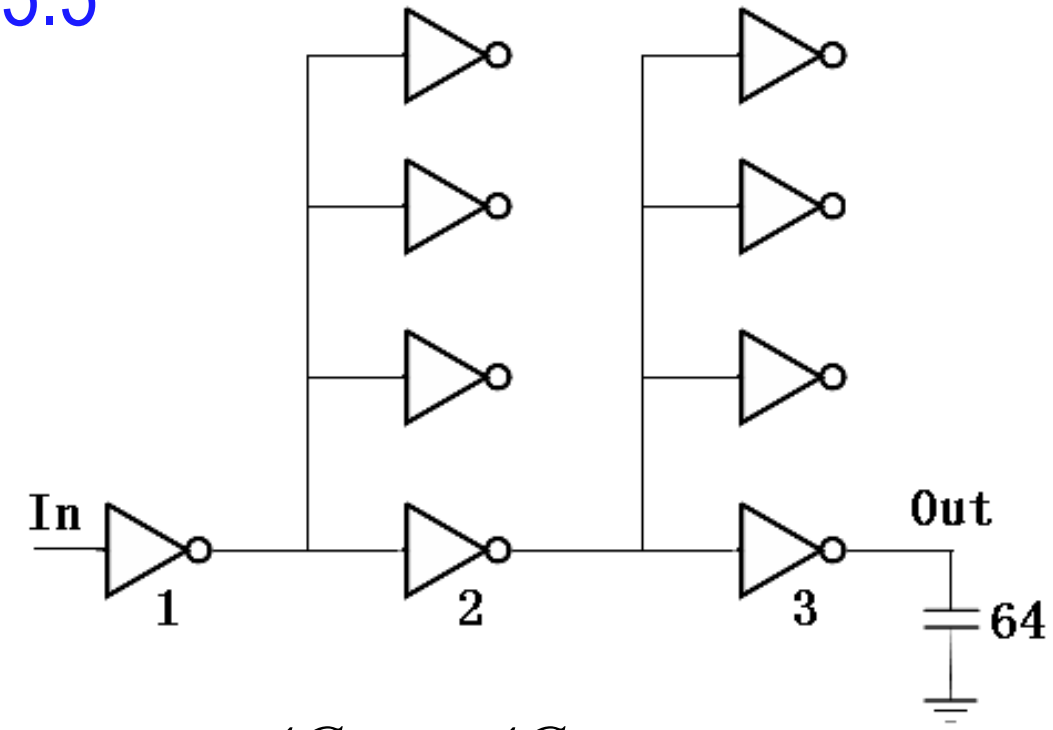
Buffer Design



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

确定反相器的尺寸

P153 例5.5



$$\frac{4C_{g2}}{C_{g1}} = \frac{4C_{g3}}{C_{g2}} = \frac{C_L}{C_{g3}}$$

$$C_{g3}=2.52C_{g2}=6.35C_{g1}$$

内容提要

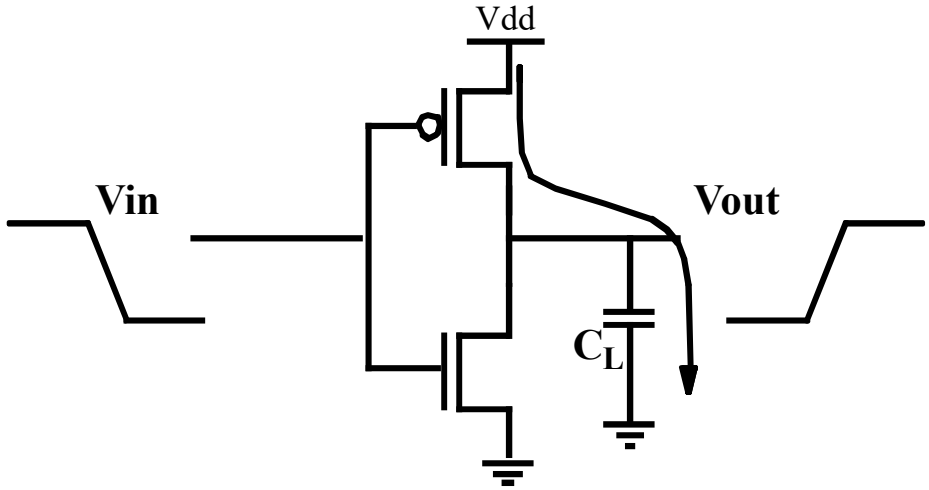
- 直观综述
- 电压传输特性（**VTC**）
- 可靠性：静态特性
- 性能：动态特性
- 功耗和能耗－延时积
- 按比例缩小技术以及对反相器的影响

5.5 CMOS中的功耗类型

- 动态功耗
  - 对电容进行充放电所消耗的能量
- 短路功耗
  - 在开关翻转期间，电源、地之间的直流电流功耗
- 漏电流功耗（静态功耗）
  - 二极管和晶体管的漏电流功耗

5.5.1 动态功耗

首先假设输入的上升和下降时间都为0，即两个晶体管不可能同时导通。



从电源中取得的能量E_{VDD}以及翻转结束时电容上存储的能量E_C为：

$$E_{VDD}=\int_0^\infty i_{VDD}(t)V_{DD}dt=V_{DD}\int_0^\infty C_L\frac{dv_{out}}{dt}dt=C_LV_{DD}\int_0^\infty dv_{out}=C_LV_{DD}^2$$

$$E_C=\int_0^\infty i_{VDD}(t)v_{out}dt=\int_0^\infty C_L\frac{dv_{out}}{dt}v_{out}dt=C_L\int_0^{¥} v_{out}dv_{out}=C_LV_{DD}^2/2$$

$$P_{dyn}=C_LV_{DD}^2f_{0\to 1}$$

$f_{0\to 1}$代表消耗能量的翻转频率（静态CMOS为$0\to 1$）。

- 能耗与晶体管尺寸无关。
- 通过减小 C_L，V_{DD}，和f来减小功耗。

工艺发展对功耗的影响

- 工作频率越来越高，即f越来越大.
- 器件密度越来越高，芯片上的总电容（C_L）也在增加。

例：0.25μm CMOS芯片，f=500MHz，平均负载电容15pF/门，扇出为4，V_{dd}=2.5V，每门功耗大约50 μW。百万门设计，50W！！

$$P=C_L V_{DD}^2 f$$

开关活动性

$$P_{dyn} = C_L V_{DD}^2 f_{0\to 1} = C_L V_{DD}^2 P_{0\to 1} f = C_{EFF} V_{DD}^2 f$$

f代表输入发生变化事件的最大概率（时钟频率）。$P_{0\to 1}$是时钟变化事件在该门的输出端引起0→1变化事件的概率。$C_{EFF}=P_{0\to 1}\, C_L$称为等效电容，它代表每个时钟周期发生开关的平均电容。

P158 例5.12

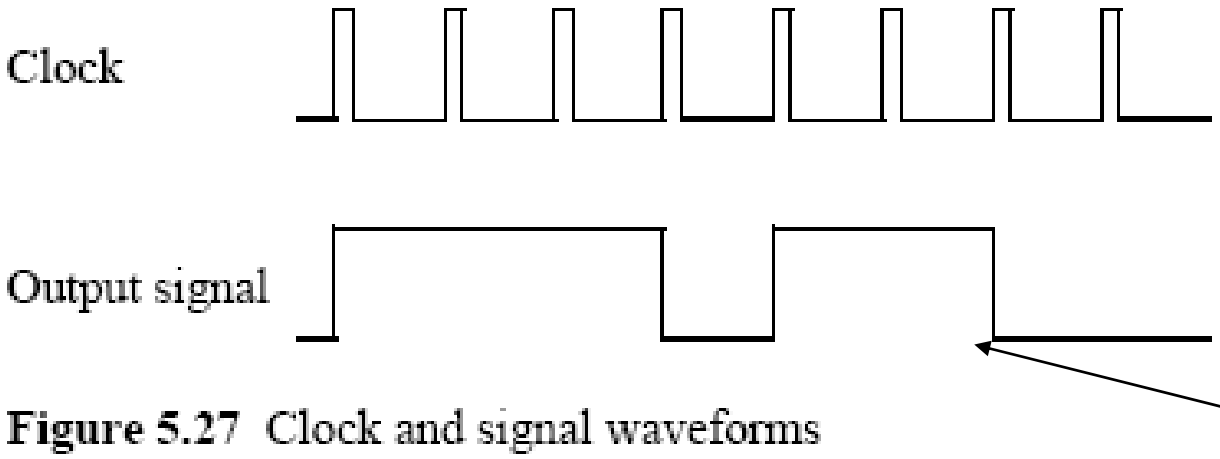
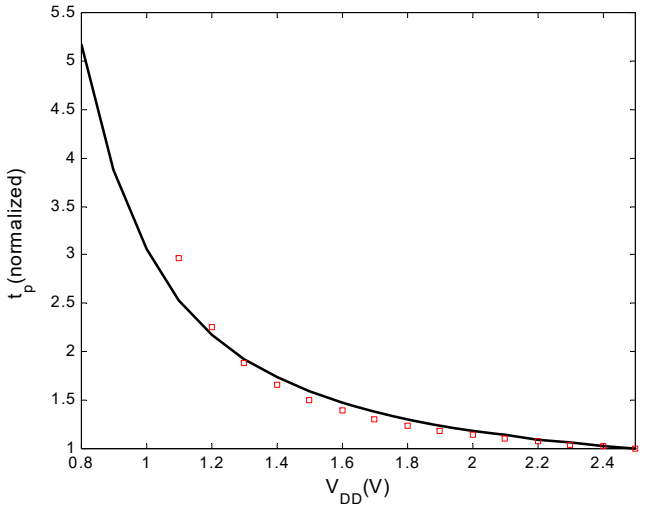


Figure 5.27 Clock and signal waveforms

0.25

减少动态功耗的探讨

- P_{dyn}正比于V_{dd}^2，降低V_{dd}（假设维持时钟频率不变）
  - V_{dd}比阈值电压高很多，没问题；
  - V_{dd}一旦接近$2V_T$，性能严重降低
- 当V_{dd}下降受限于性能时，
  只能减小等效电容
  减小实际电容和翻转活动性



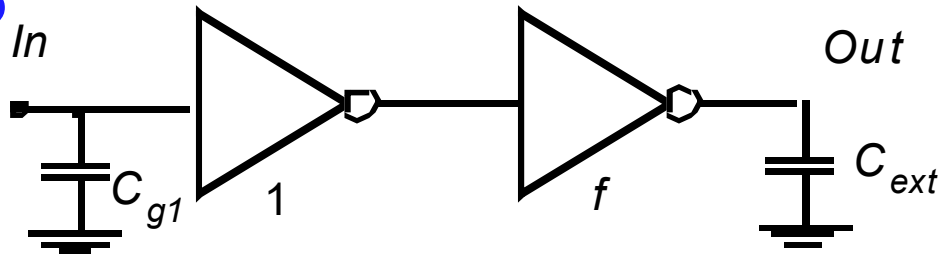
在逻辑和结构的抽象层次上实现

有利于改善电路的性能

实际上，保持最小尺寸，会影响电路的性能，通过逻辑或结构上的加速来解决

Transistor Sizing for Minimum Energy

P158 例5.13



- Goal: Minimize Energy of whole circuit
  - Design parameters: f and V_{DD}
  - $t_p \le t_{pref}$ of circuit with f=1 and $V_{DD}=V_{ref}$

$$t_p = t_{p0}\left(\left(1+\frac{f}{\gamma}\right)+\left(1+\frac{F}{f\gamma}\right)\right)$$

$$t_{p0} \propto \frac{V_{DD}}{V_{DD}-V_{TE}}$$

$V_{TE}=V_T+V_{DSAT}/2$

由公式（5.21)推导出的近似表达式

Transistor Sizing (2)

- 性能约束 (γ=1)

即尺寸放大器的传播延时应当等于（或小于）参考电路（*f*=1，$V_{dd}=V_{ref}$）的延时。假设该门的本征输出电容等于它的栅电容，即γ=1。

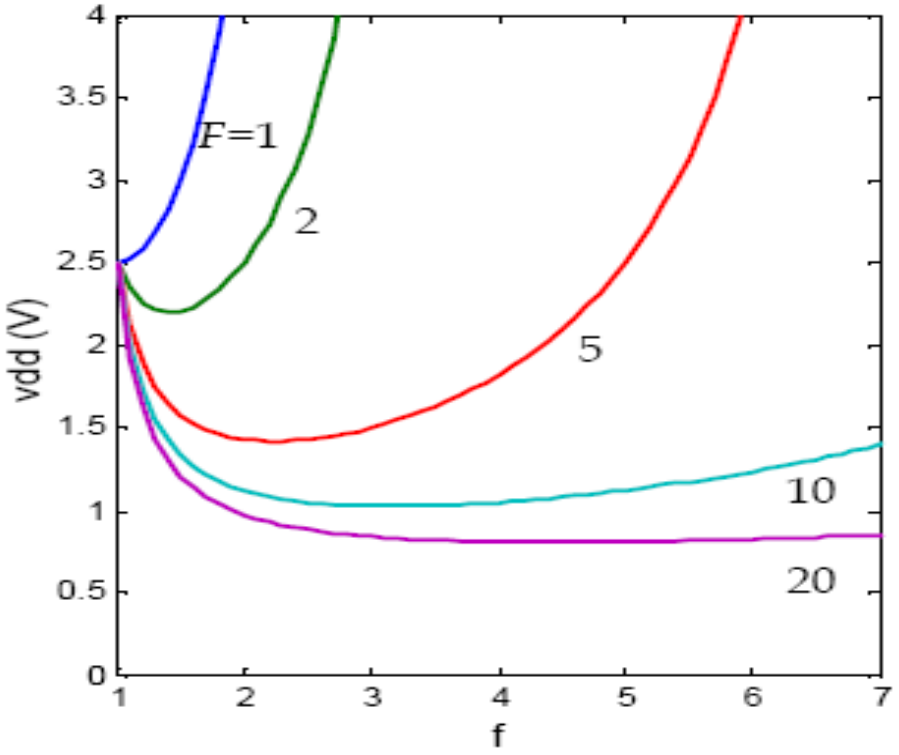
$$\frac{t_p}{t_{pref}} = \frac{t_{p0}}{t_{p0ref}} \frac{\left(2+f+\frac{F}{f}\right)}{(3+F)} = \frac{V_{DD}}{V_{ref}} \frac{V_{ref}-V_{TE}}{V_{DD}-V_{TE}} \frac{\left(2+f+\frac{F}{f}\right)}{(3+F)} = 1$$

建立了尺寸系数f和电源电压之间的关系。

尺寸系数f与电源电压之间的需求关系

V_{DD}=f(*f*)

- 对于不同的F值，这些曲线都有一个明显的最小值。从最小尺寸起增加反相器的尺寸最初会使性能提高，因此允许降低电源电压。这在达到最优尺寸f=$F^{1/2}$之前都是有效的。进一步加大器件尺寸只会增加自载系数而降低性能，因此需要提高电源电压。



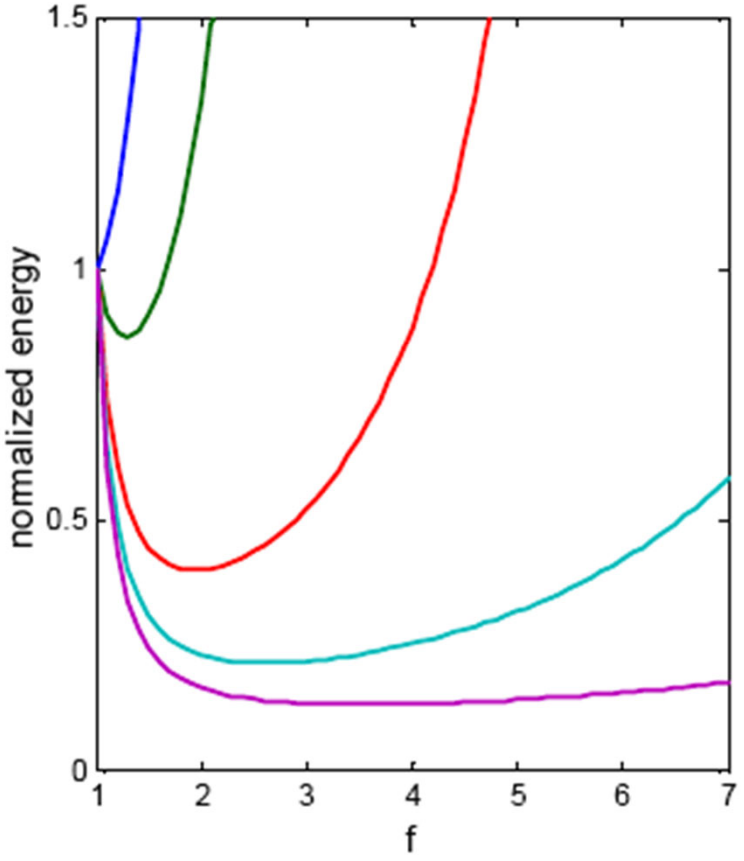
Energy for single Transition

输入端单个翻转的能耗为：

$$E = V_{DD}^2 C_{g1}\left[(1+\gamma)(1+f)+F\right]$$

$$\frac{E}{E_{ref}} = \left(\frac{V_{DD}}{V_{ref}}\right)^2 \left(\frac{2+2f+F}{4+F}\right)$$

E/E_{ref}=f(f)



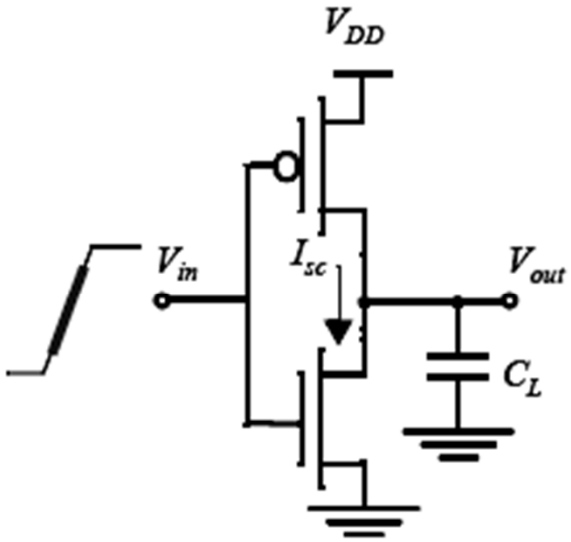
- 改变器件尺寸并降低电源电压是减小逻辑电路能耗的有效方法。
- 在最优值之外过多的加大晶体管尺寸会付出较大的能量代价。
- 考虑能量时的最优尺寸系数小于考虑性能时的最优尺寸系数，在F较大时尤其如此。例如当扇出为20时，f_{opt}（能量）＝3.53；而f_{opt}（性能）＝4.47。一旦Vdd开始接近V_{TE}，加大器件尺寸只能很少地降低电压，因此能耗的降低也很少。

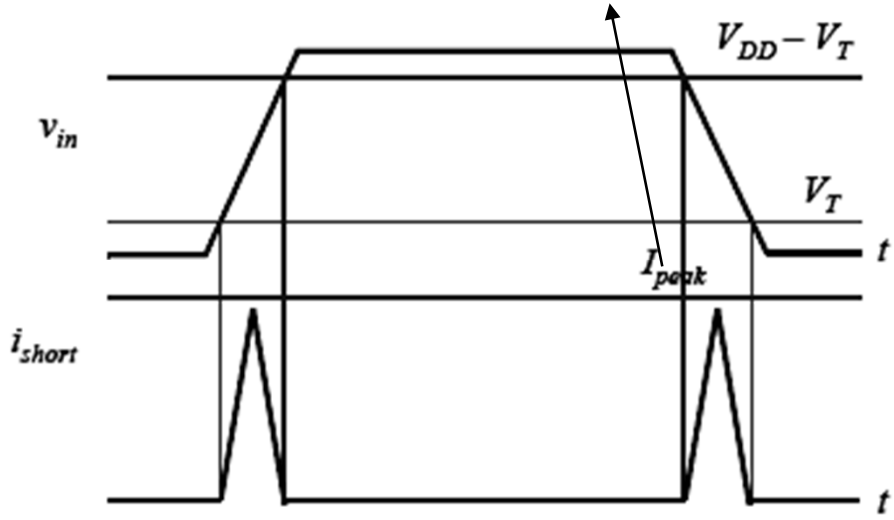
短路功耗

- 在实际设计中，输入波形的上升和下降时间不为零。
- 输入信号斜率不为无穷大造成了在开关过程中在V_{DD}和V_{SS}之间在短期内出现一条直接通路，此时两个MOS管同时导通，导致能量的消耗。

Short Circuit Currents

由器件的饱和电流决定，因此正比于晶体管的尺寸





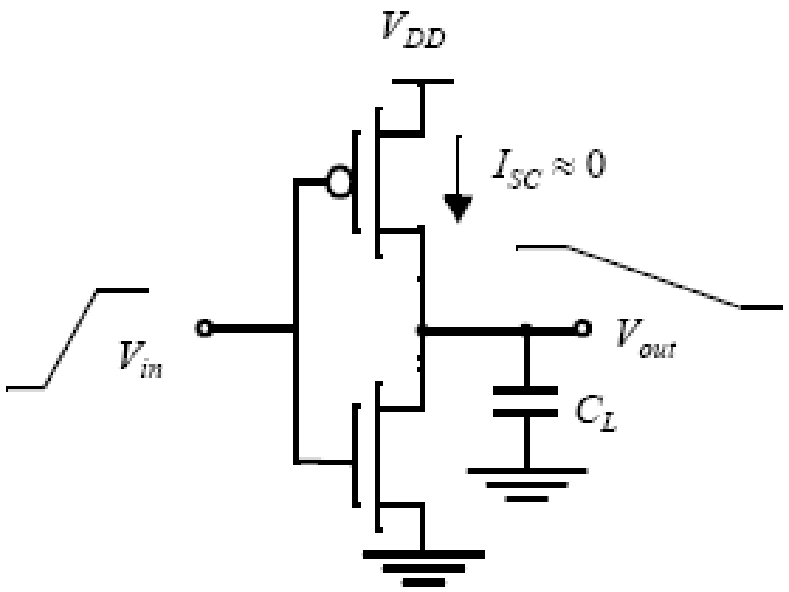
每个开关周期消耗的能量：$$E_{dp} = V_{DD}\frac{I_{peak}t_{sc}}{2} + V_{DD}\frac{I_{peak}t_{sc}}{2} = t_{sc}V_{DD}I_{peak}$$

平均功耗：$$P_{dp} = t_{sc}V_{DD}I_{peak}f = C_{sc}V_{DD}^2 f$$

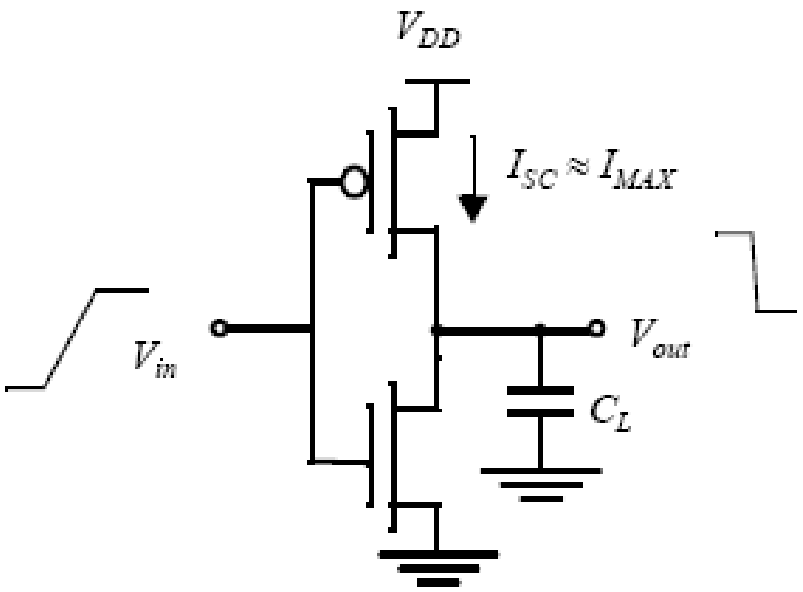
两个器件同时导通的时间：$$t_{sc} = \frac{V_{DD}-2V_T}{V_{DD}}t_s \approx \frac{V_{DD}-2V_T}{V_{DD}} \times \frac{t_{r(f)}}{0.8}$$

t_s:0～100%的翻转时间

负载电容对短路电流的影响



(a) Large capacitive load

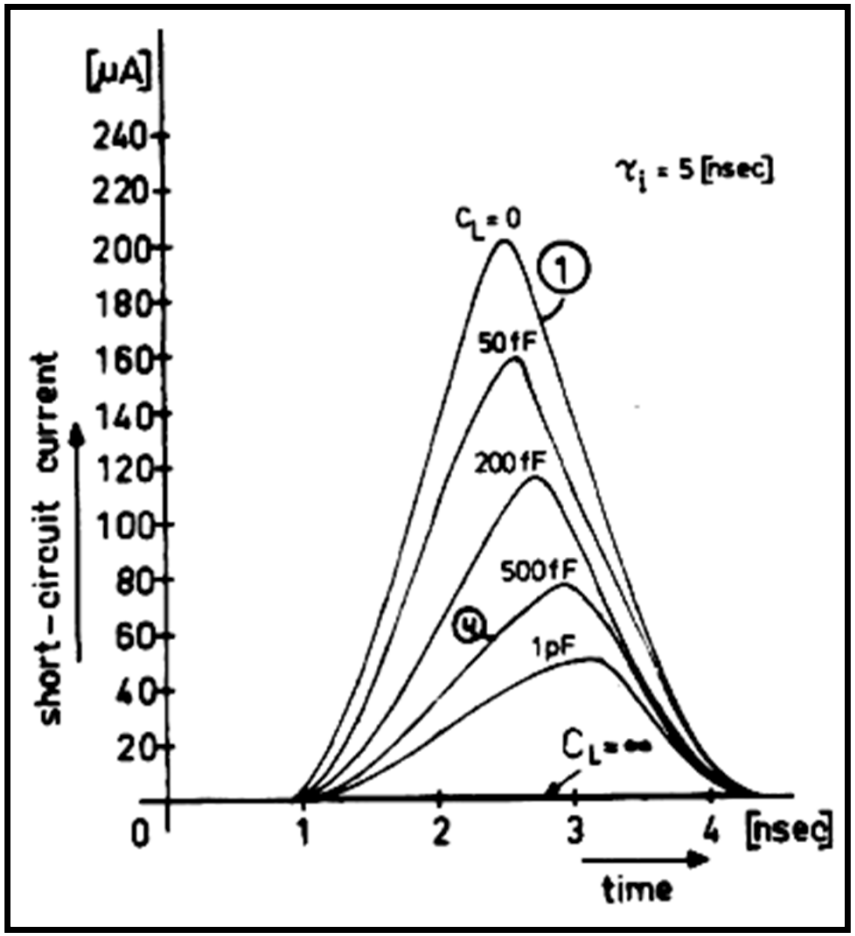


(b) Small capacitive load

输入在输出开始改变之前就已经通过了过渡区，这一时期**PMOS**的源—漏电压近似为**0**，该器件还没有传导任何电流就断开了。

PMOS器件的源—漏电压在翻转期间的大部分时间内等于**Vdd**，从而引起了最大的短路电流（等于**PMOS**的饱和电流）。

How to keep Short-Circuit Currents Low?



结论：使输出的上升/下降时间大于输入的下降/上升时间可以使短路功耗减到最小。

但输出的上升/下降时间太大会降低电路的速度；并在扇出门中引起短路电流。

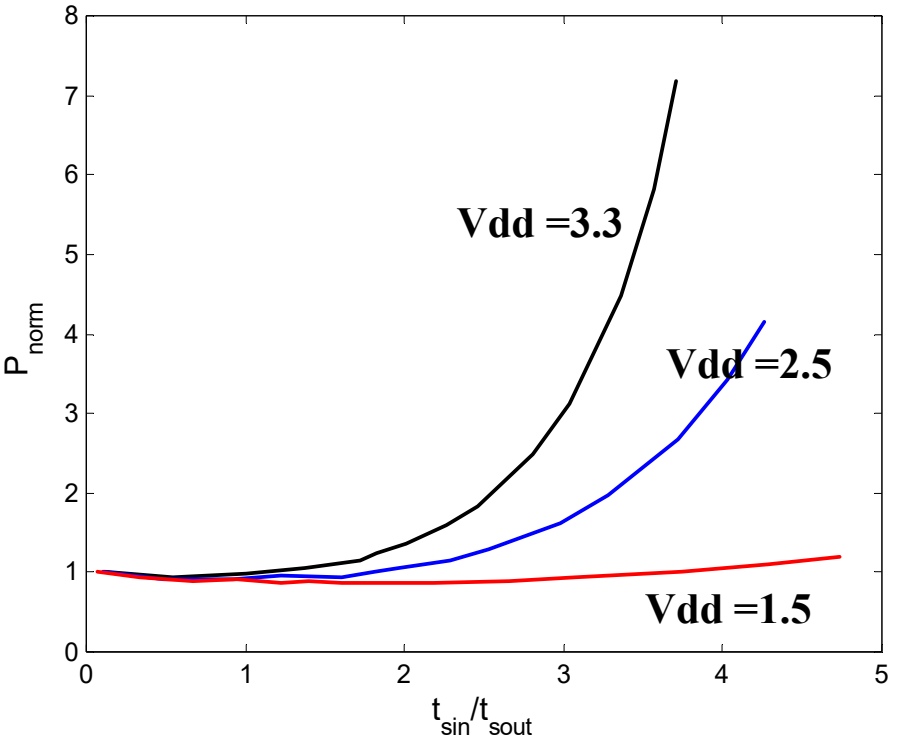
只顾局部优化而不管全局是会引起不良后果的。

Minimizing Short-Circuit Power

对于一个给定的反相器尺寸，当负载电容太小时，功耗主要来自短路电流；对于非常大的负载电容值，所有的功耗都来自对负载的充电和放电。

如果输入和输出的上升/下降时间相等，则大部分功耗与动态功耗有关，只有很小一部分（**<10%**）来自短路电流。

当降低电源电压时，短路电流减小。当阈值电压以比电源电压低的速率下降时，短路功耗在深亚微米工艺中变得不重要。



5.5.2 静态功耗

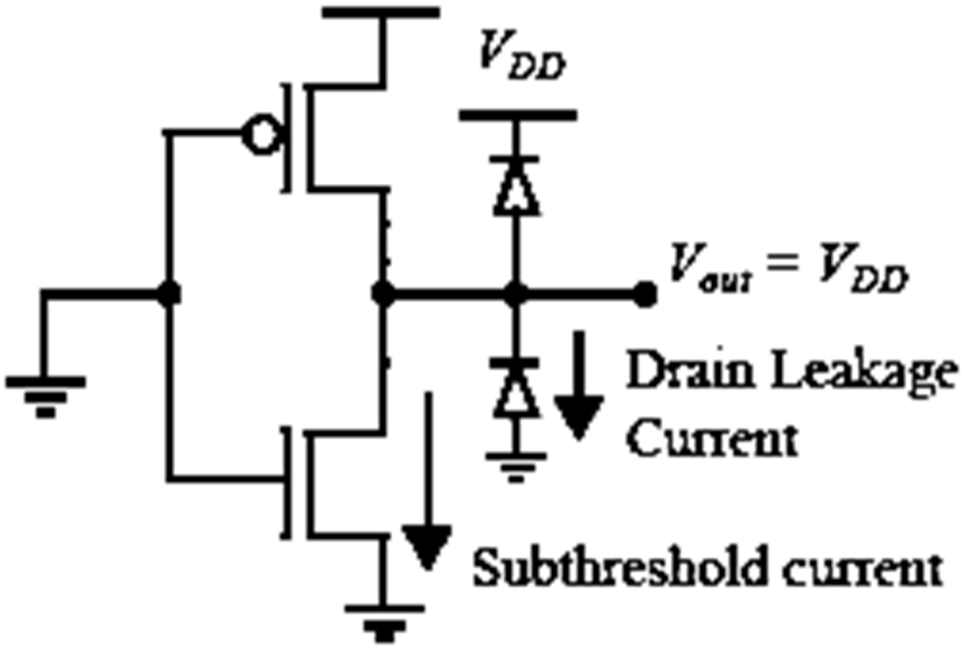
- 一个电路处于静态（或稳态）时的功耗

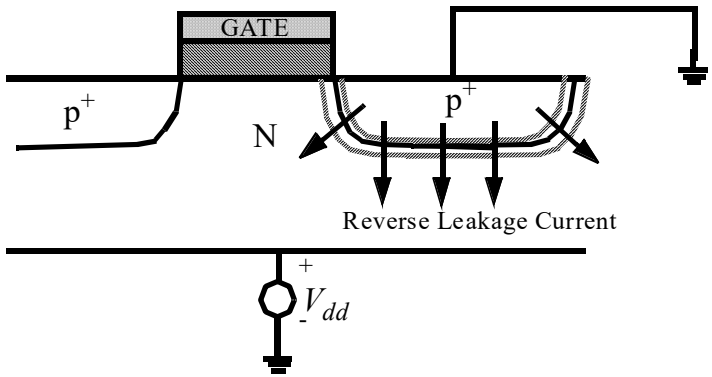
$$P_{stat} = I_{stat} V_{DD}$$

- 理想情况下，**CMOS**反相器的静态电流为零，因为两个晶体管在稳态工作状况下决不会同时导通。
- 总会有泄漏电流流过位于晶体管源（或漏）与衬底之间的反相偏置的二极管结。

泄漏电流

一般来说，泄漏电流非常小，可以忽略不计.





$$I_{DL} = J_S \times A$$

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

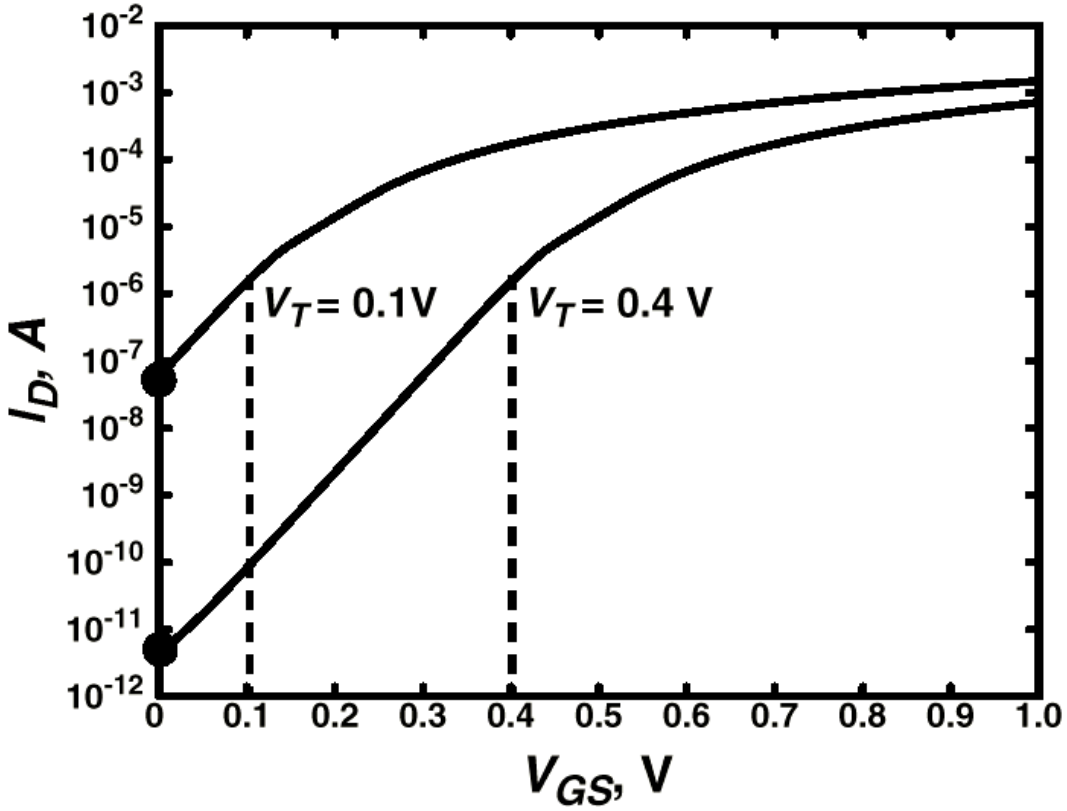
亚阈值电流是超低功耗电路设计中面临的最大问题！

泄漏电流的本质

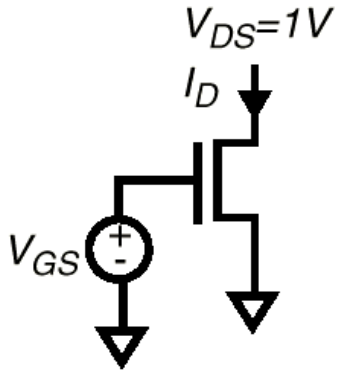
- 是由热产生的载流子引起的
- 数值随结上的温度按指数关系增加
  - 例如，85摄氏度时，漏电流为室温时的60倍，因此总是期望电路工作温度较低。
- 一个越来越突出的来源是晶体管的亚阈值电流
  - 阈值电压越是接近0V，则在$V_{GS}=0$时漏电流越大，因而静态功耗也就越大。
  - 器件的阈值电压一般应保持足够高。标准工艺的特征值V_T从未小于0.5~0.6V。

Subthreshold Leakage Component

$$I_D \sim I_s e^{\frac{qV_{GS}}{nkT}}, \quad n = 1 + \frac{C_D}{C_{ox}}$$



60mV/10倍电流



- **Leakage control is critical for low-voltage operation**

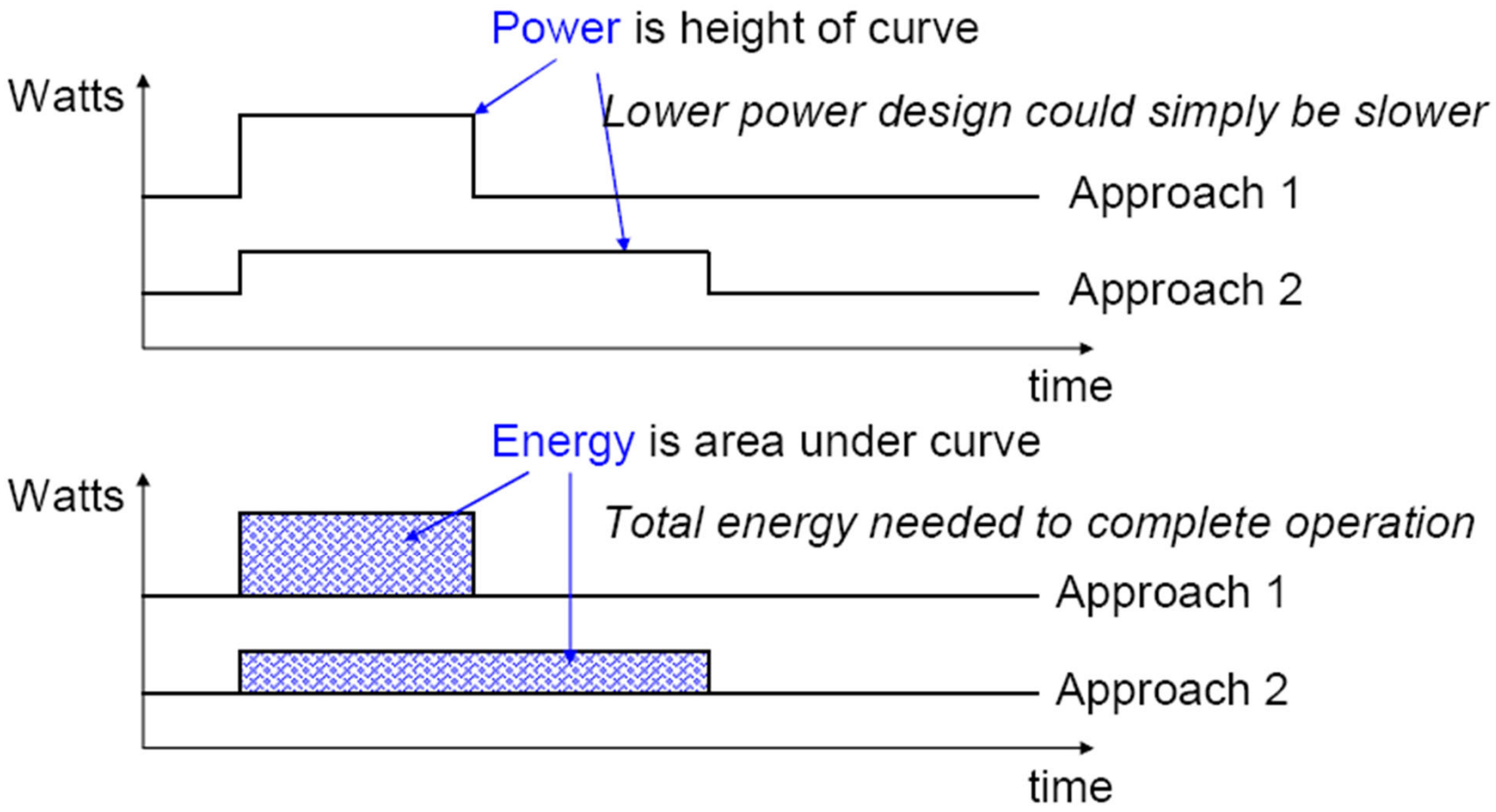
工艺发展对漏电流的影响

- 工艺尺寸的缩小，出现了电源电压的降低。
- 降低电源电压同时阈值电压不变会造成性能严重的损失，可以降低器件的阈值电压解决这一问题。
- 降低阈值电压虽然可以避免电源电压降低的损失，但是阈值电压的最低值由所允许的亚阈值漏电流决定。
- 电源电压的继续降低预示着新一代**CMOS**工艺的出现，但它也迫使阈值电压更为降低，从而使亚阈值导电成为功耗的主要来源。
- 生产具有迅速彻底关断特性的器件的工艺技术——**SOI**技术。

功耗和能量

- 消耗的功耗（瓦特W）
  - 决定电池的寿命
  - 决定封装的要求
- 峰值功耗（Peak Power）
  - 决定电源、地线的设计
  - 影响信号的噪声容限和可靠性分析
- 能量的效率（焦耳J）
  - 能量消耗的速率
  - 能量＝功耗×延时

功耗与能量


Power is height of curve
Watts
Lower power design could simply be slower
Approach 1
Approach 2
time
Energy is area under curve
Watts
Total energy needed to complete operation
Approach 1
Approach 2
time

功耗—延时积

- 功耗CMOS反相器的总功耗

$$P_{tot}=P_{dyn}+P_{dp}+P_{stat}=(C_L V_{DD}^2+V_{DD}I_{peak}t_s)f_{0\to 1}+V_{DD}I_{leak}$$

- 功耗-延时积

$$(PDP) = P_{av} * t_p$$

假设这个门以其最大可能的速率$f_{max}=1/(2t_p)$切换，并忽略静态和短路功耗。

$$PDP=C_L V_{DD}^2 f_{max} t_p = C_L V_{DD}^2/2$$

PDP衡量了开关这个门所需要的能量。但对于给定的结构PDP可以通过降低电源电压减小。但这个电路工作的最优电压应当是仍然能够保证功能的最低可能的电压，但这样牺牲了性能。因此需要把性能和能量一起考虑。

能量—延时积

能量-延时积

把性能和能量的度量放在一起考虑。

$$(EDP) = PDP * t_p$$

$$EDP = PDP \times t_p = P_{av} t_p^2 = \frac{C_L V_{DD}^2}{2} t_p$$

假设NMOS和PMOS具有可比拟的阈值电压和饱和电压，简化公式5.21得到：

$$t_{pHL} = 0.69 \frac{3}{4} \frac{C_L V_{DD}}{I_{DSATn}} = 0.52 \frac{C_L V_{DD}}{(W/L)_n k'_n V_{DSATn}(V_{DD} - V_{Tn} - V_{DSATn}/2)} \rightarrow t_p \approx \frac{\alpha C_L V_{DD}}{V_{DD} - V_{Te}}$$

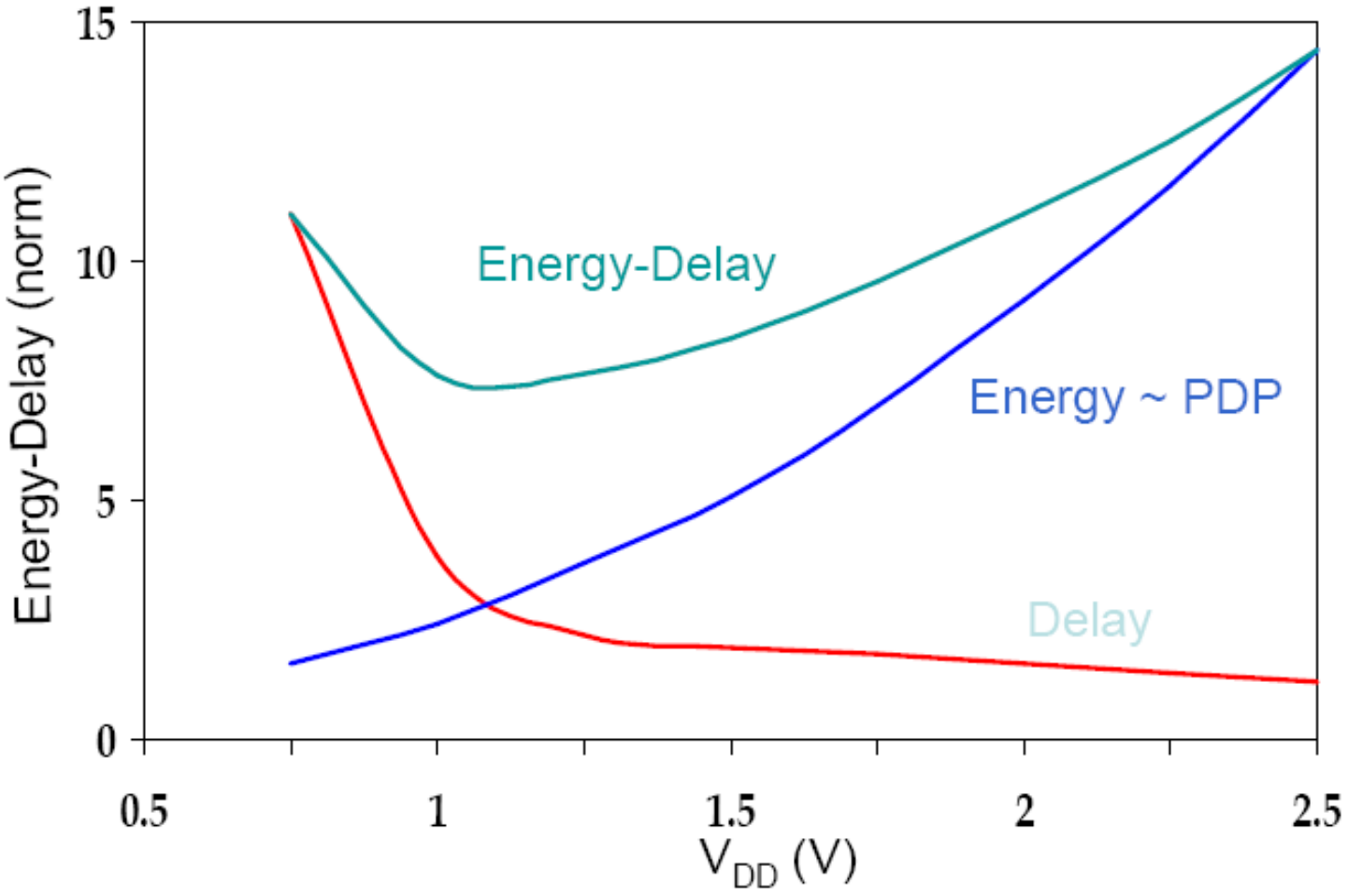
$$EDP = \frac{\alpha C_L^2 V_{DD}^3}{2(V_{DD} - V_{TE})}$$

对V_{dd}求导并令结果为0

$$V_{DDopt} = \frac{3}{2} V_{TE}$$

EDP作为V_{DD}的函数

P165 例5.15



Principles for Power Reduction

- 常用方法: 降低电压!
  - 近几年来电源电压的降低是实现低功耗的主要手段。
  - 极低电压的应用会带来新的问题 **(0.6 … 0.9 V)**
- 降低开关活动性
- 减小寄生电容
  - **Device Sizing: for *F*=20**
    - **f_{opt}(energy)=3.53, f_{opt}(performance)=4.47**

内容提要

- 直观综述
- 电压传输特性（**VTC**）
- 可靠性：静态特性
- 性能：动态特性
- 功耗和能耗－延时积
- 按比例缩小技术以及对反相器的影响

工艺缩小的目标

- 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

工艺缩小的内容

- 几何尺寸
  - 前端：晶体管几何尺寸**W/L**，氧化层厚度
  - 后端：互连，例如线、接触孔、…
- 阈值电压
- 电源电压

Technology Scaling—宏观上的影响

- 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 Evolution (2000 data)

International Technology Roadmap for Semiconductors

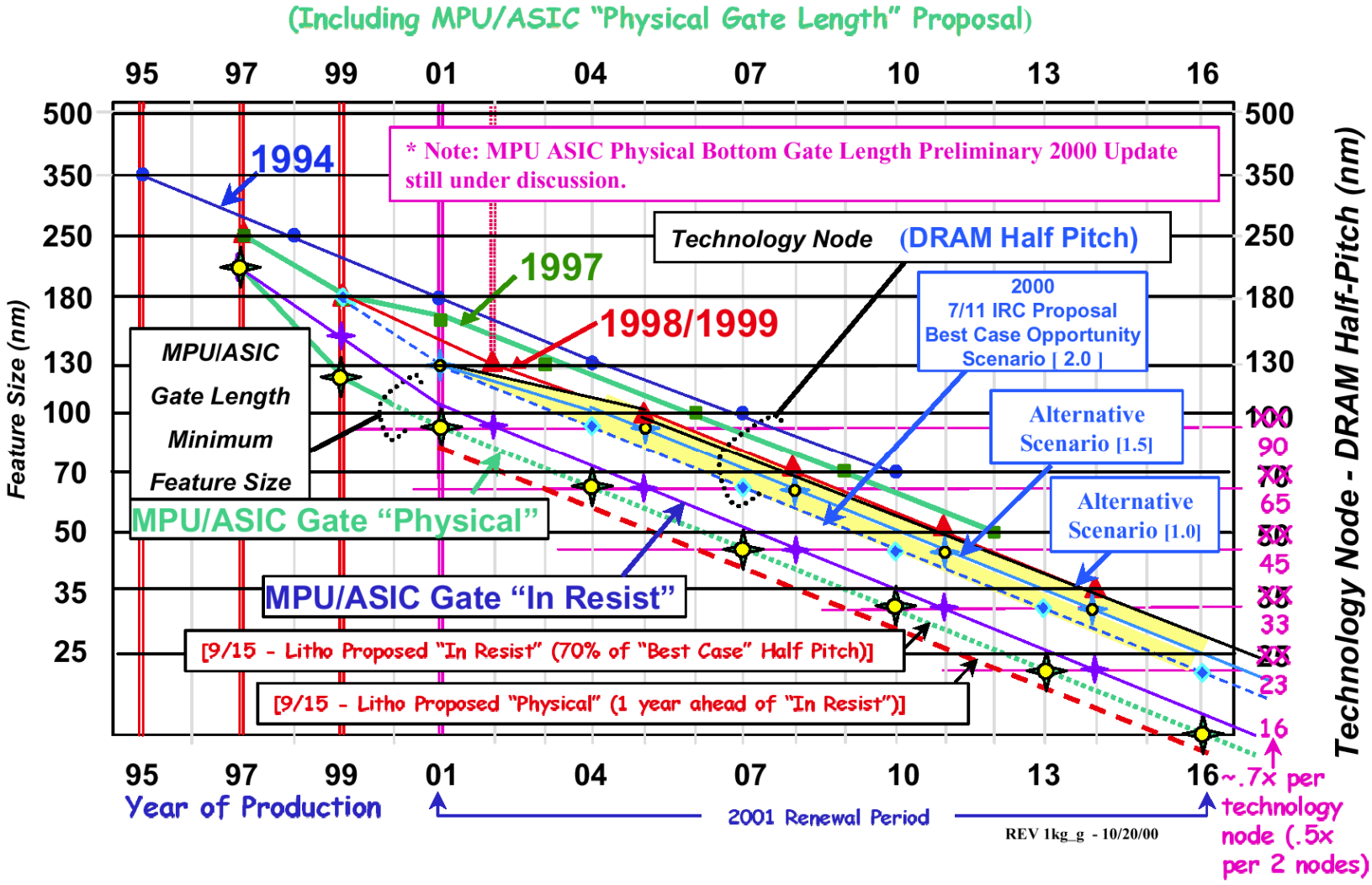
Year of Introduction	1999	2000	2001	2004	2008	2011	2014
Technology node [nm]	180		130	90	60	40	*30*
Supply [V]	1.5-1.8	1.5-1.8	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6	*0.3-0.6*
Wiring levels	6-7	6-7	7	8	9	9-10	*10*
Max frequency [GHz],Local-Global	1.2	1.6-1.4	2.1-1.6	3.5-2	7.1-2.5	11-3	*14.9 -3.6*
Max μP power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

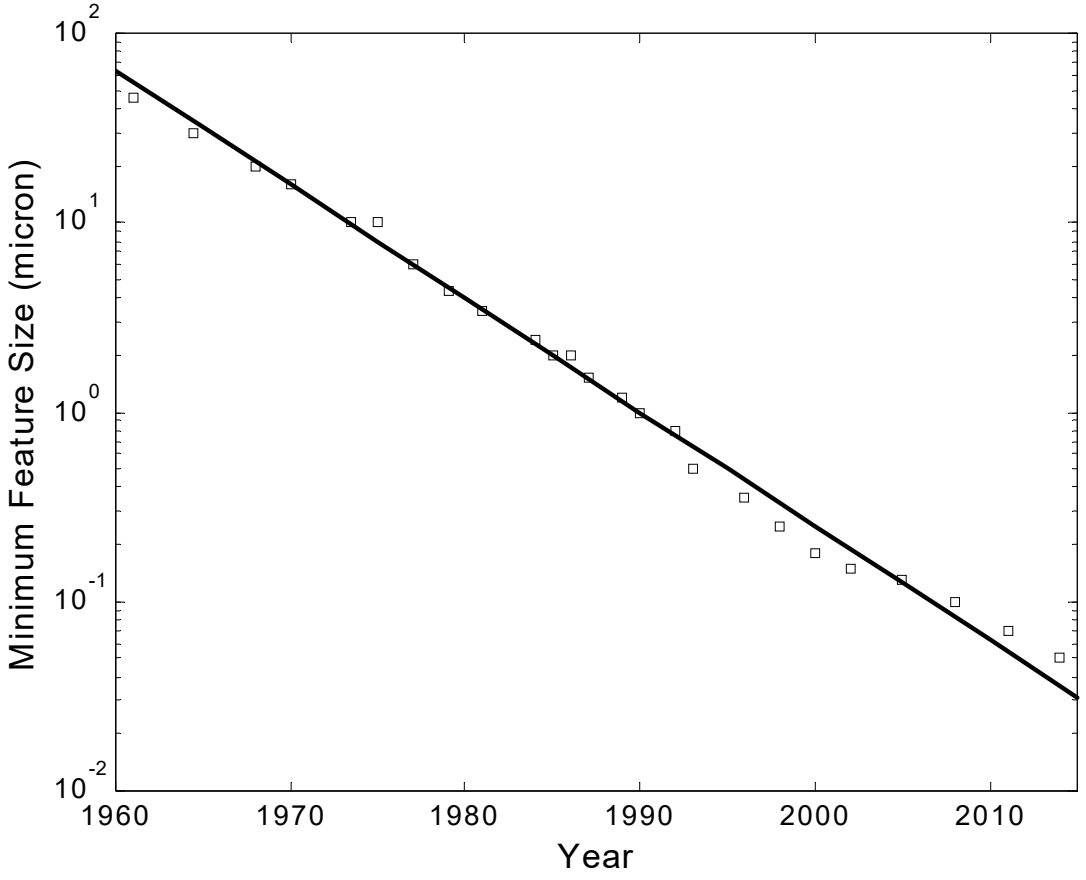
Technology Evolution (1999)

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

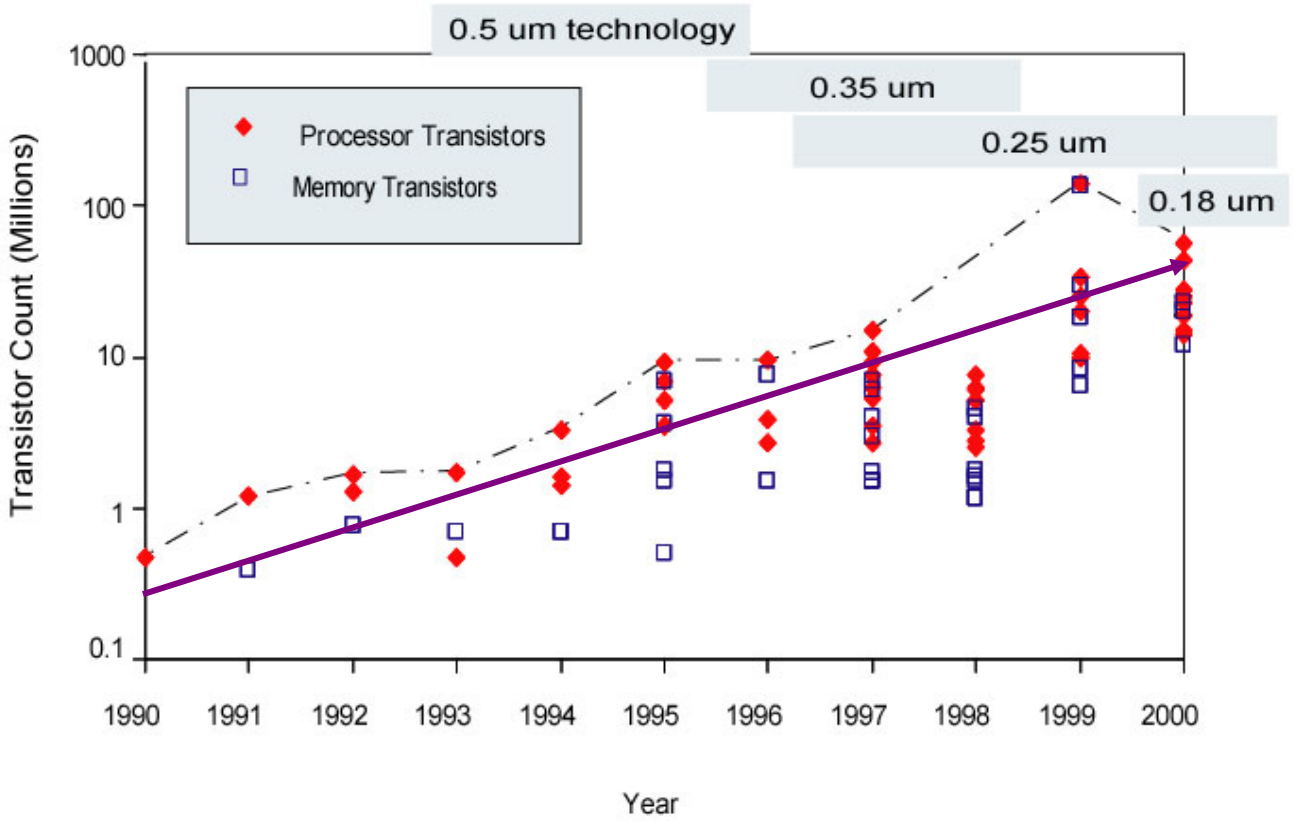


Technology Scaling (1)



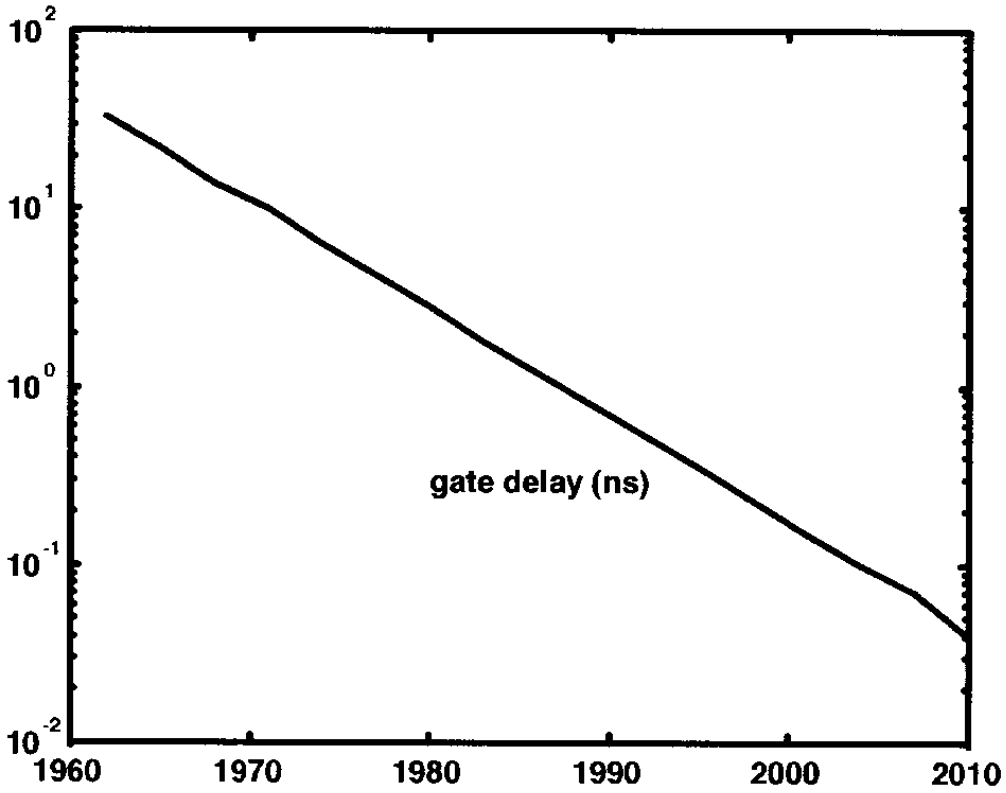
Minimum Feature Size

Technology Scaling (2)



Number of components per chip

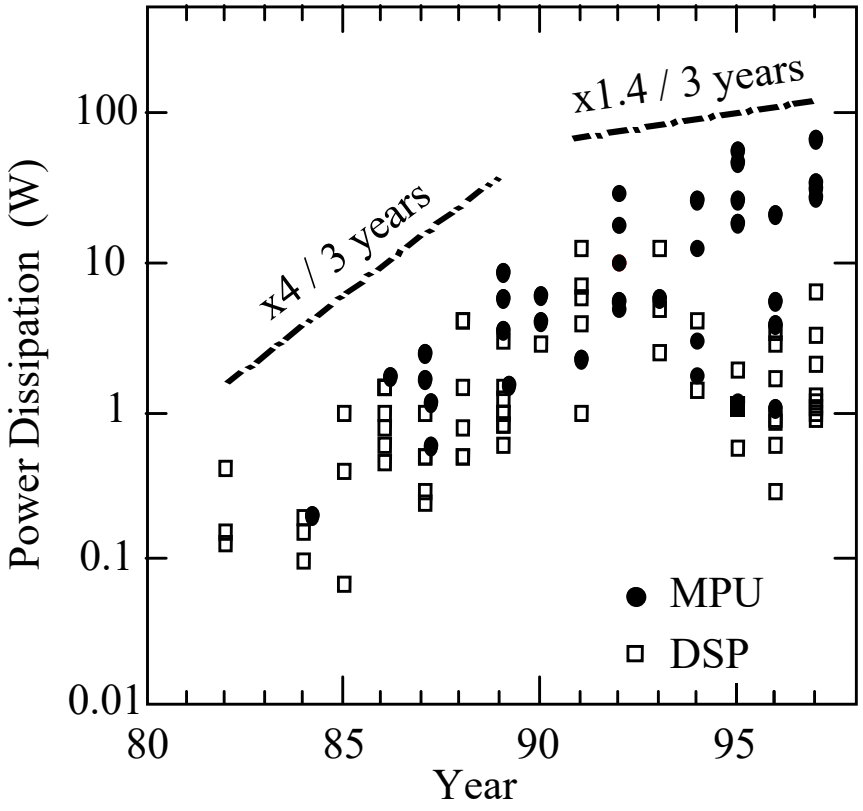
Technology Scaling (3)



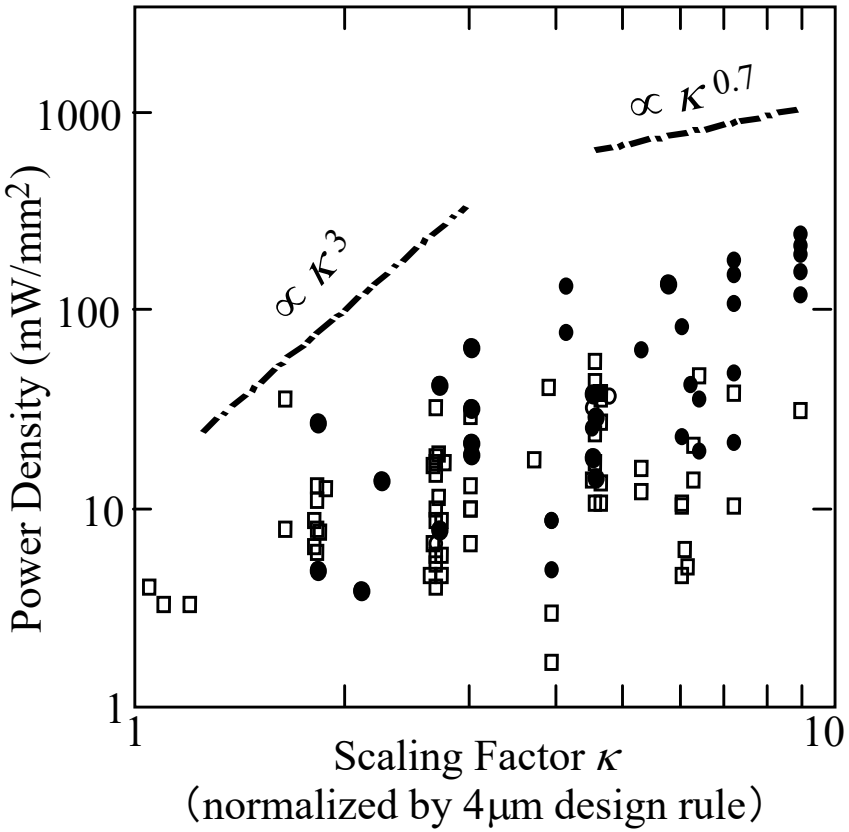
Propagation Delay

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

Technology Scaling (4)



(a) Power dissipation vs. year.



(b) Power density vs. scaling factor.

From Kuroda

缩放规则（I）

定义

- 所有器件的尺寸都缩小同一个因子S（S>1代表尺寸缩小），包括晶体管宽度和长度、栅氧厚度及结深。
- 所有的电压缩放同一个比例U，包括阈值、电源电压。

全比例缩小（恒电场缩小）

- S＝U。
- 保持缩小器件中电场强度不变：V/L。
- 理想情况下：提高器件密度（减小面积）、提高性能（减小本征延时）、降低功耗。

缩放规则（II）

- 恒压缩小
  - S可变，U＝1。
  - 出于兼容性的原因，电压不可能随意缩放。在20世纪90年代之前，5V一直是标准电压，3.3V、2.5V只是在0.5μm工艺出现之后才确定地位的。
  - 当今，由于速度饱和问题，器件尺寸和电压缩小的比例相近。
  - 在速度饱和情况下，如果只缩小器件尺寸而不降低电压，对功耗和性能产生不良影响。

缩放规则（III）

- 一般化缩小
  - 器件尺寸和电压都缩小，但缩小比例不同
  - S>U>1

  电源电压并不像工艺尺寸的缩小那么快，当工艺尺寸从0.5um缩小到0.1um时，最大的电源电压从5V降低到1.5V。
- 一般化缩小的原因：
  - 一些本征的器件电压（如硅的带隙电压和内建结电势）是材料参数，因此不能缩小。
  - 晶体管阈值电压的缩小潜力是有限的。阈值电压太低将使完全关断器件非常困难。这在阈值有比较大的工艺偏差时尤为严重。

Scaling Relationships （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_{depl}^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_{av}	$k_{n,p}V^2$	1/S	S/U^2	S
t_p (intrinsic)	C_LV / I_{av}	1/S	U/S^2	$1/S^2$
P_{av}	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
W, L, t_{ox}		$1/S$	$1/S$	$1/S$
V_{DD}, V_T		$1/S$	$1/U$	1
N_{SUB}	V/W_{depl}^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_{gate}	$C_{ox}WL$	$1/S$	$1/S$	$1/S$
k_n, k_p	$C_{ox}W/L$	S	S	S
I_{sat}	$C_{ox}WV$	$1/S$	$1/U$	1
Current Density	$I_{sat}/Area$	S	S^2/U	S^2
Ron	V/I_{sat}	1	1	1
Intrinsic Delay	$R_{on}C_{gate}$	$1/S$	$1/S$	$1/S$
P	$I_{sat}V$	$1/S^2$	$1/U^2$	1
Power Density	$P/Area$	1	S^2/U^2	S^2