



中国科学技术大学

University of Science and Technology of China

数字集成电路设计

第五章 功 耗

白雪飞

中国科学技术大学微电子学院

- 引言
- 动态功耗
- 静态功耗
- 能量-延时优化
- 低功耗体系结构



引言

■ 瞬时功率 (Instantaneous Power)

- 电路元件消耗或提供的瞬时功率定义为该元件的电流和电压的乘积

$$P(t) = I(t)V(t)$$

■ 能量 (Energy)

- 在某一时间间隔 T 内消耗或提供的能量是瞬时功率的积分

$$E = \int_0^T P(t) dt$$

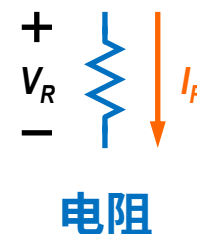
■ 平均功率 (Average Power)

- 在某一时间间隔 T 上的平均功率定义为

$$P_{\text{avg}} = \frac{E}{T} = \frac{1}{T} \int_0^T P(t) dt$$

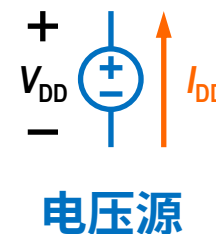
■ 电阻消耗的瞬时功率

$$P_R(t) = \frac{V_R^2(t)}{R} = I_R^2(t) R$$



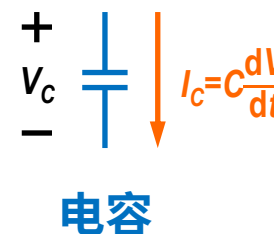
■ 电压源提供的瞬时功率

$$P_{V_{DD}}(t) = I_{DD}(t) V_{DD}$$



■ 电容充电/放电时存储/释放的能量

$$\begin{aligned} E_C &= \int_0^{\infty} I(t) V(t) dt = \int_0^{\infty} C \frac{dV}{dt} V(t) dt \\ &= C \int_0^{V_C} V(t) dV = \frac{1}{2} C V_C^2 \end{aligned}$$



CMOS反相器翻转过程



■ 当输入从“1”翻转到“0”时

- NMOS管截止而PMOS管导通，将负载电容充电至 V_{DD}
- 存储在负载电容中的能量为

$$E_C = \frac{1}{2} C_L V_{DD}^2$$

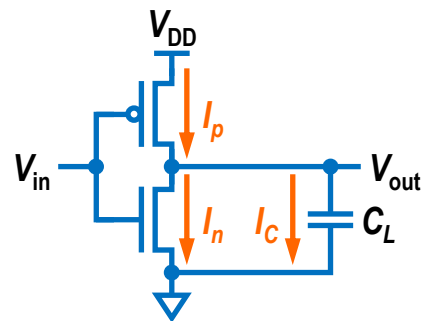
- 电源提供的能量为

$$E_{V_{DD}} = \int_0^{\infty} I(t) V_{DD} dt = \int_0^{\infty} C_L \frac{dV}{dt} V_{DD} dt = C_L V_{DD} \int_0^{\infty} dV = C_L V_{DD}^2$$

- 电源提供的能量中，一半存储在负载电容中，另一半消耗在PMOS管中

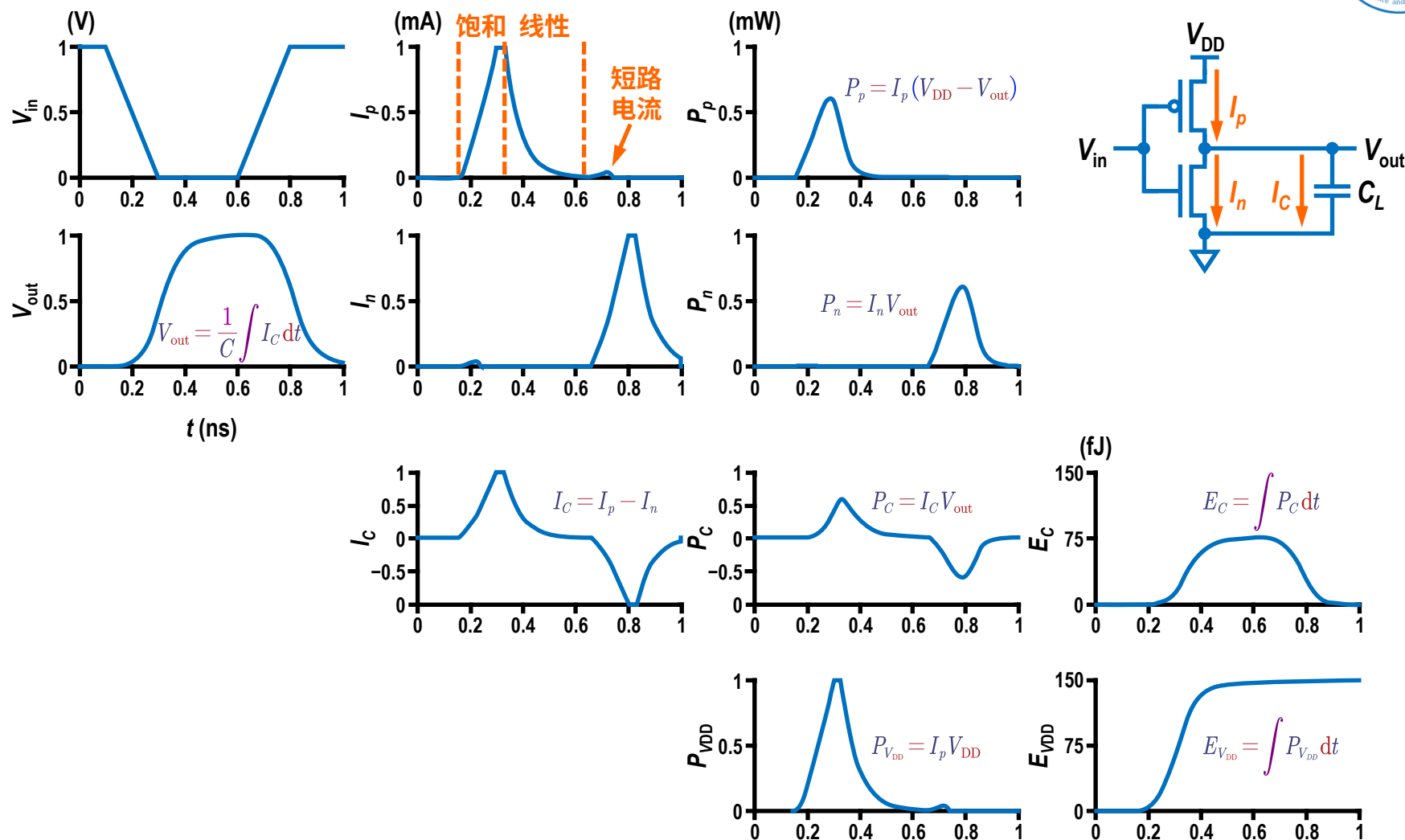
■ 当输入从“0”翻转回“1”时

- PMOS管截止而NMOS管导通，存储在电容中的能量释放并消耗在NMOS管中
- 这一翻转跳变过程中没有从电源获取任何能量



CMOS反相器及负载电容

CMOS反相器翻转过程



CMOS反相器翻转过程中电压、电流、功率和能量变化

电源电压：1.0 V，输入信号频率：1 GHz，负载电容：150 fF

■ 翻转功耗

- 假设逻辑门以平均频率 f_{sw} 翻转，在时间间隔 T 内，负载电容将被充电和放电 Tf_{sw} 次，则平均翻转功耗为

$$P_{\text{switching}} = \frac{E}{T} = \frac{Tf_{sw}CV_{DD}^2}{T} = CV_{DD}^2 f_{sw}$$

- 大多数逻辑门并非在每个时钟周期都发生翻转，平均翻转功耗也可以表示为

$$P_{\text{switching}} = \alpha CV_{DD}^2 f$$

- 时钟频率 f
- 活动因子(Activity Factor) α
 - 节点从0跳变至1的概率
 - 时钟的活动因子为1，静态CMOS逻辑的活动因子经验值约为0.1

■ 短路功耗

- 晶体管翻转过程中，上拉网络和下拉网络同时部分导通造成的短路电流功耗

■ 动态功耗 (Dynamic Power)

- 翻转功耗 (Switching Power)
- 短路电流 (Short Current)

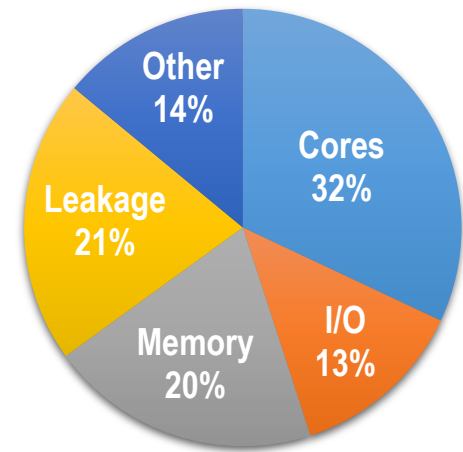
■ 静态功耗 (Static Power)

- 亚阈值泄漏电流 (Subthreshold Leakage)
- 栅泄漏电流 (Gate Leakage)
- 结泄漏电流 (Junction Leakage)
- 有比电路中的竞争电流 (Contention Current)

$$P_{\text{dynamic}} = P_{\text{switching}} + P_{\text{shortcircuit}}$$

$$P_{\text{static}} = (I_{\text{sub}} + I_{\text{gate}} + I_{\text{junct}} + I_{\text{contention}})V_{\text{DD}}$$

$$P_{\text{total}} = P_{\text{dynamic}} + P_{\text{static}}$$



Niagara2处理器的功耗

65-nm CMOS, 84 W@1.1 V, 1.4 GHz
(JSSC, 43(1), 2008, 6–20)

动态功耗

■ 翻转功耗

- 动态功耗大部分由翻转功耗构成
- 节点的电容是此节点上栅电容、扩散电容、连线电容之和
- 节点的等效电容是其实际电容与活动因子之积
- 在电源电压和频率已知的情况下，翻转功耗取决于所有节点的等效电容之和

■ 短路功耗

- 通常小于整个功耗的10%，可以比较保守地估计为翻转功耗的10%
- 纳米工艺下的短路功耗通常可以忽略不计

■ 动态功耗的优化方法

- 选择能满足目标性能的最低工作频率
- 选择能支持目标工作频率的最低电源电压
- 通过使不需要工作的模块进入休眠状态减小活动因子
- 通过优化电路减小每一部分的总负载电容

动态功耗估算举例



- **例：**某1-V电源电压65-nm工艺的片上数字系统，沟道长度为50 nm，即 $\lambda=25$ nm。芯片共有 10^9 个晶体管，其中：
逻辑管：数量50 M，平均宽度 12λ ，平均活动因子0.1；
存储管：数量950 M，平均宽度 4λ ，平均活动因子0.02。
晶体管的栅电容为1 fF/ μm ，扩散电容为0.8 fF/ μm ，忽略连线电容。

- **求：**试估算芯片工作在1 GHz时的翻转功耗。

- **解：**

逻辑管总电容 $C_{\text{logic}} = (50 \times 10^6) (12\lambda) (0.025 \mu\text{m}/\lambda) ((1 + 0.8) \text{ fF}/\mu\text{m}) = 27 \text{ nF}$

存储管总电容 $C_{\text{mem}} = (950 \times 10^6) (4\lambda) (0.025 \mu\text{m}/\lambda) ((1 + 0.8) \text{ fF}/\mu\text{m}) = 171 \text{ nF}$

翻转功耗 $P_{\text{switching}} = (0.1 \times C_{\text{logic}} + 0.02 \times C_{\text{mem}}) (1.0 \text{ V})^2 (10^9 \text{ Hz}) = 6.12 \text{ W}$

■ 降低活动因子

- 降低活动因子是降低动态功耗的非常有效和易于实现的途径
- 若电路完全关断，则其活动因子和动态功耗都降为零
- 时钟门控 (Clock Gating)：通过停止时钟来关断电路模块

■ 活动因子的估算

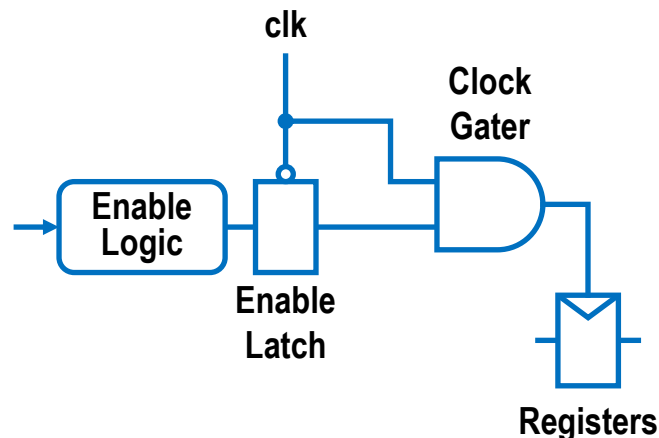
- 时钟信号的活动因子为1
- 真随机数据的活动因子为0.25
- 静态CMOS逻辑的活动因子经验值约为0.1
- 逻辑门的活动因子可以通过计算其翻转概率进行估算
- 毛刺(Glitch)会增加活动因子

■ 时钟门控的作用

- 将时钟信号与使能信号相“与”来关断闲置电路模块的时钟
- 有效降低活动因子和节点电容
 - 时钟信号的活动因子非常高, $\alpha=1$
 - 时钟网络具有很大的节点电容
 - 关断寄存器时钟可以阻止其翻转, 并停止其下游组合逻辑的翻转活动
- 需要判断电路模块是否为闲置状态

■ 时钟门控逻辑电路

- 使能信号可直接用于带使能端的寄存器
- 时钟有效时, 使能信号必须保持稳定
 - 对于上升沿触发的系统, 可使用锁存器保证使能信号只在时钟低电平期间变化
- 时钟门控可用于电路模块时钟网络前端



时钟门控逻辑

■ 翻转概率

- 节点的活动因子是其从“0”翻转到“1”的概率
- 与电路的逻辑功能有关
- 由节点为逻辑“1”的概率可以估算活动因子
- 毛刺会引起额外的翻转，并提高活动因子

■ 活动因子的估算

- 令 P_i 为节点 i 处于逻辑“1”的概率，则节点 i 的活动因子为

$$\alpha_i = \overline{P_i} P_i; \quad \overline{P_i} = 1 - P_i$$

- 逻辑门的输出概率可由输入概率计算得到
- 当路径包含重聚的扇出时，输入信号之间存在相关性，因而需要应用条件概率计算
- 没有更好数据的情况下，可估计为 $\alpha=0.1$

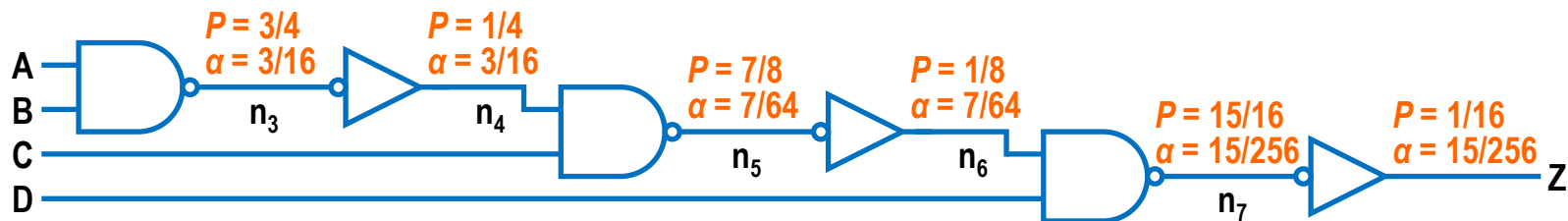
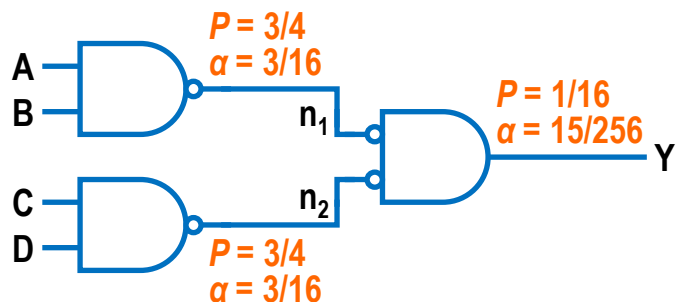
逻辑门输出逻辑“1”的概率

逻辑门	P_Y
AND2	$P_A P_B$
AND3	$P_A P_B P_C$
OR2	$1 - \overline{P_A} \overline{P_B}$
NAND2	$1 - P_A P_B$
NOR2	$\overline{P_A} \overline{P_B}$
XOR2	$P_A \overline{P_B} + \overline{P_A} P_B$

翻转概率举例

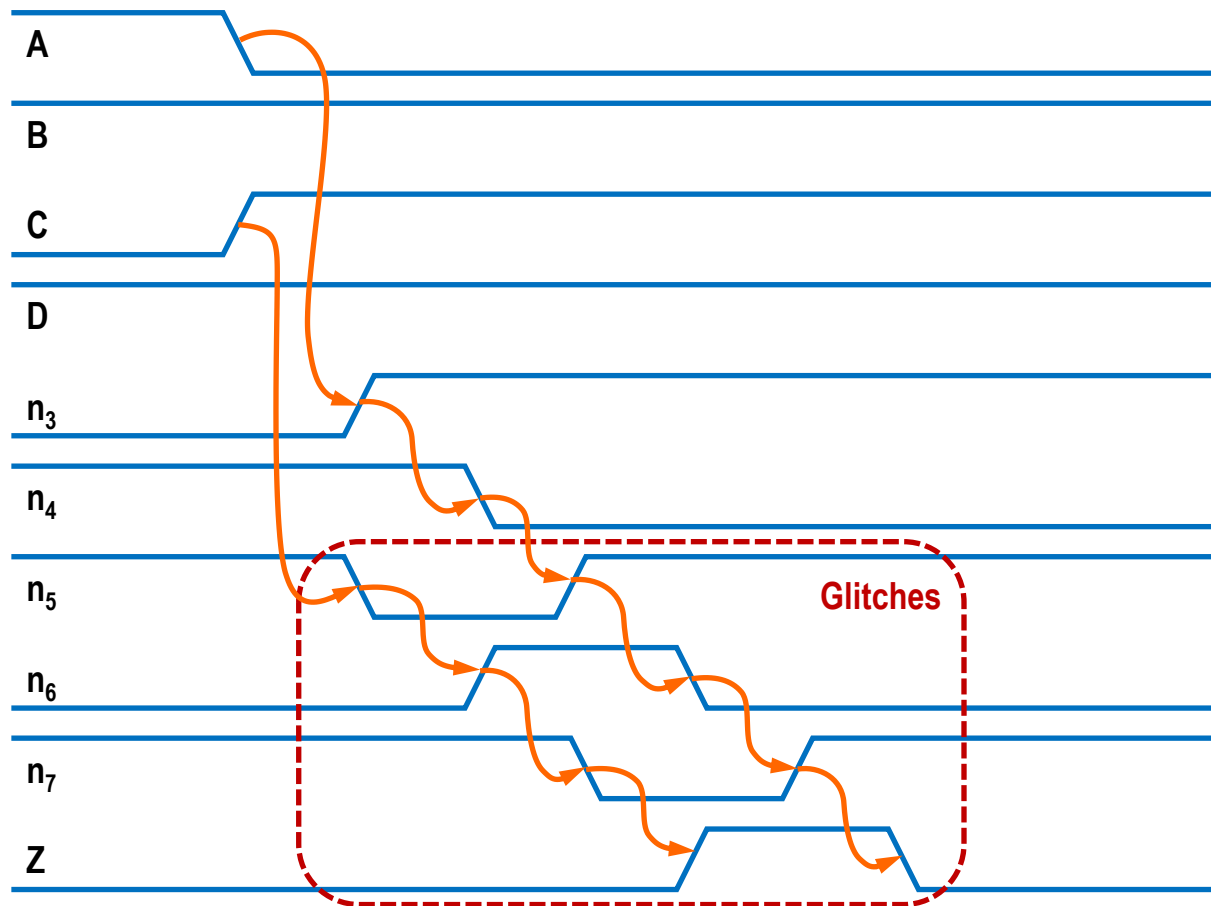


- 例：若所有输入信号为逻辑“1”的概率皆为 $P=0.5$ ，求以下四输入与门电路中各节点的活动因子。



四输入与门电路及节点信号概率和活动因子

毛刺 (Glitch)



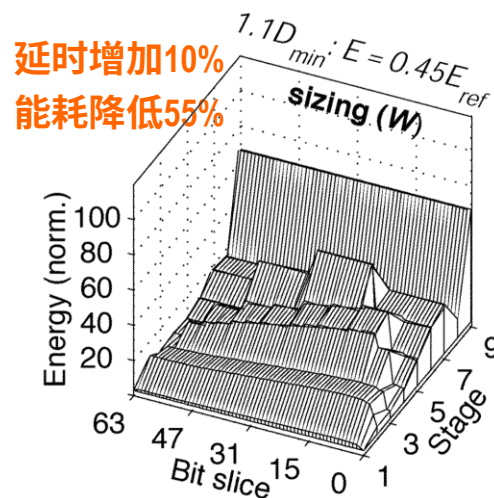
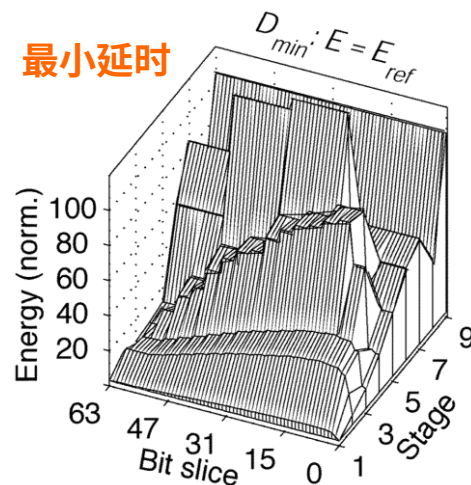
逻辑门链中的毛刺

■ 翻转电容

- 来自电路中的连线和晶体管
- 良好的平面规划和布局可以降低连线电容
- 选择较少的逻辑级数和较小的晶体管可以降低器件的电容

■ 晶体管尺寸选择

- 非关键路径上采用最小尺寸的门
- 采用较大的每级努力，仅比最小延时稍微增加一些，就可以显著减小晶体管尺寸，节省很大比例的能量
- 缩小具有较高活动因子或较大尺寸的门
- 采用反相器或缓冲器驱动长连线，而不采用具有较高逻辑努力的复杂门



在延时约束下调整门的尺寸

(JSSC, 39(8), 2004, 1282-1293)

■ 电压和频率的选择

- 每个电路模块都运行在能够满足性能需求的最低电压和最低频率
- 芯片划分成多个电压域，或根据工作模式调整电源电压
- 芯片划分成多个频率域，降低频率可以减小晶体管尺寸或降低电源电压

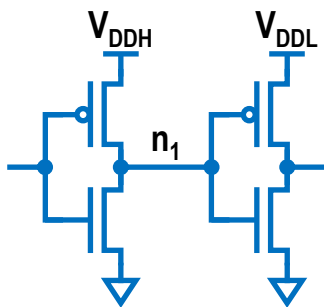
■ 电压域 (Voltage Domain)

- 为不同电路模块分别提供不同的电源电压
- 每个电压域根据电路的时序和特性需要进行电源电压优化
- 当信号从低电压域传输至高电压域时，需要使用电平转换器

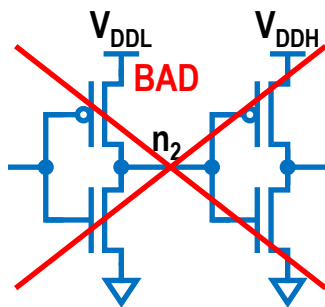
■ 动态电压/频率调整 (Dynamic Voltage/Frequency Scaling, DVFS)

- 根据工作负荷动态调整电源电压和时钟频率

当 n_1 为高电平 V_{DDH} 时， V_{DDL} 域的门翻转更快，需注意污染延时变化

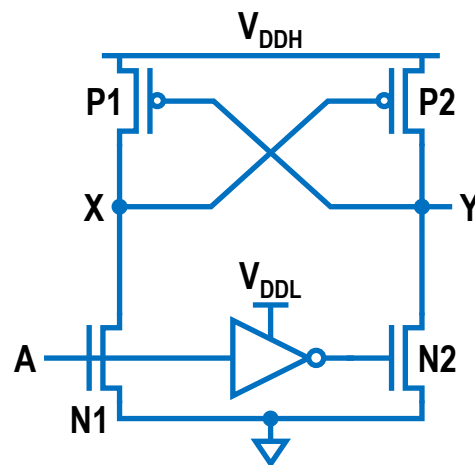


当 n_2 为高电平 V_{DDL} 时， V_{DDH} 域PMOS管 $V_{gs} < 0$ ，导通或增加泄漏电流



高低电压域的反相器直连

A=0时，N2导通，Y=0，X=1 (V_{DDH})
A=1时，N1导通，X=0，Y=1 (V_{DDH})

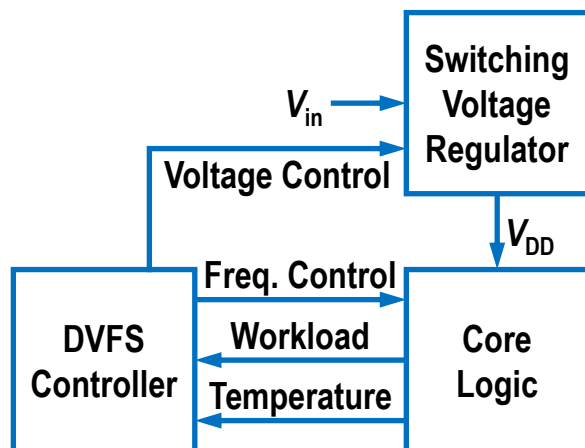


电平转换器

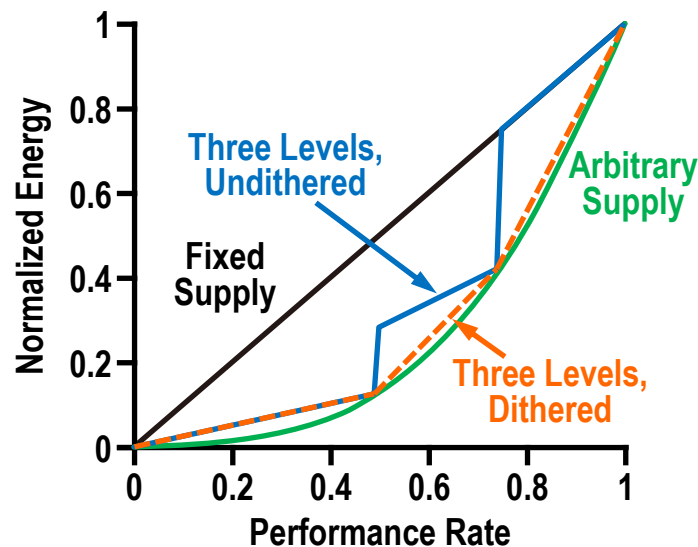
动态电压/频率调整



- 动态电压调整 (Dynamic Voltage Scaling, DVS)
- 动态电压/频率调整 (Dynamic Voltage/Frequency Scaling, DVFS)
- 超动态电压调整 (Ultra-Dynamic Voltage Scaling, UDVS)



DVFS系统



DVFS降低能耗

静态功耗

■ 静态功耗

- 即使在芯片处于静态(Quiescent)时也存在功耗
- 具有低阈值电压和薄栅氧的纳米工艺中，静态功耗约占总功耗的1/3

■ 静态功耗来源

- 泄漏：名义上关断的器件中流过的电流
 - 亚阈值泄漏：截止晶体管的沟道
 - 栅泄漏：栅极电容
 - 结泄漏：源/漏扩散区与衬底之间的反偏二极管
- 竞争电流：有比电路中导通晶体管之间流过的电流

■ 亚阈值泄漏电流

- 应当截止的晶体管中流过的漏源电流

$$I_{ds} = I_{\text{off}} \cdot 10^{\frac{V_{gs} + \eta(V_{ds} - V_{DD}) - k_{\gamma} V_{sb}}{S}} \left(1 - e^{\frac{-V_{ds}}{v_T}} \right)$$

- 当 $V_{ds} > 50 \text{ mV}$ 时，简化为

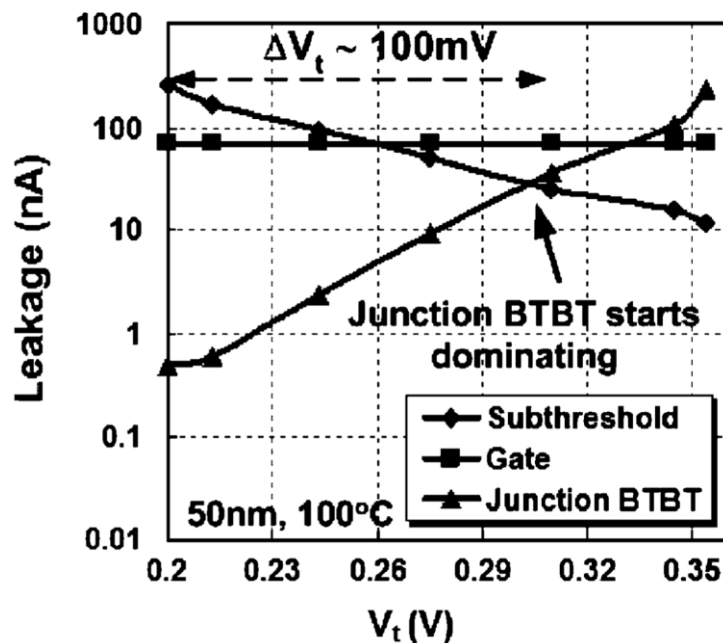
$$I_{\text{sub}} = I_{\text{off}} \cdot 10^{\frac{V_{gs} + \eta(V_{ds} - V_{DD}) - k_{\gamma} V_{sb}}{S}}$$

■ 各项参数在65-nm工艺的典型取值

- $V_{gs}=0$ 且 $V_{ds}=V_{DD}$ 时的亚阈值电流

$$I_{\text{off}} = \begin{cases} 100 \text{ nA}/\mu\text{m} @ V_t = 0.3 \text{ V} \\ 10 \text{ nA}/\mu\text{m} @ V_t = 0.4 \text{ V} \\ 1 \text{ nA}/\mu\text{m} @ V_t = 0.5 \text{ V} \end{cases}$$

- DIBL系数 $\eta = 100 \text{ mV/V}$
- 体效应系数 $k_{\gamma} = 0.1$
- 亚阈值斜率 $S = 100 \text{ mV/decade}$



泄漏与阈值电压的关系

(TVLSI, 15(6), 2007, 660–671)

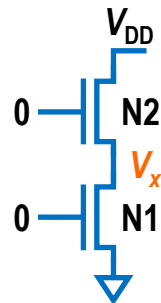
■ 串联截止晶体管的泄漏电流

- 如右图串联截止晶体管，假设 $V_x > 50 \text{ mV}$ ，则有

$$I_{\text{sub}} = \underbrace{I_{\text{off}} \cdot 10^{\frac{\eta(V_x - V_{\text{DD}})}{S}}}_{\text{N1}} = \underbrace{I_{\text{off}} \cdot 10^{\frac{-V_x + \eta((V_{\text{DD}} - V_x) - V_{\text{DD}}) - k_\gamma V_x}{S}}}_{\text{N2}}$$

$$V_x = \frac{\eta V_{\text{DD}}}{1 + 2\eta + k_\gamma}$$

$$I_{\text{sub}} = I_{\text{off}} \cdot 10^{\frac{-\eta V_{\text{DD}} \left(\frac{1 + \eta + k_\gamma}{1 + 2\eta + k_\gamma} \right)}{S}} \approx I_{\text{off}} \cdot 10^{\frac{-\eta V_{\text{DD}}}{S}}$$



串联截止晶体管

■ 堆叠效应 (Stack Effect)

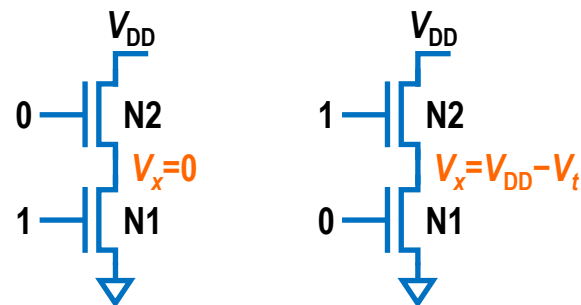
- 串联截止晶体管的泄漏电流显著减低
- 两个截止晶体管堆叠，亚阈值泄漏降低至约 $1/10$ ($V_{\text{DD}} = 1.0 \text{ V}$)
- 三个或更多截止晶体管堆叠，亚阈值泄漏更低

■ 栅泄漏

- 当电压应用于栅极时载流子隧穿薄栅介质引起
- 栅泄漏与介质厚度和栅源电压有极强的相关性
- PMOS管栅泄漏比NMOS管低一个数量级，可以忽略不计

■ 串联堆叠晶体管的栅泄漏

- 如右图串联堆叠晶体管
- 若N1导通而N2截止，则
 - N1: $V_{gs}=V_{DD}$ ，栅泄漏最大
 - N2: 截止，无栅泄漏
- 若N1截止而N2导通，则
 - N1: 截止，无栅泄漏
 - N2: $V_{gs}=V_t$ ，栅泄漏可忽略不计
- 使晶体管堆叠并使截止晶体管靠近电源/地可以降低栅泄漏



串联堆叠晶体管

亚阈值泄漏和栅泄漏举例



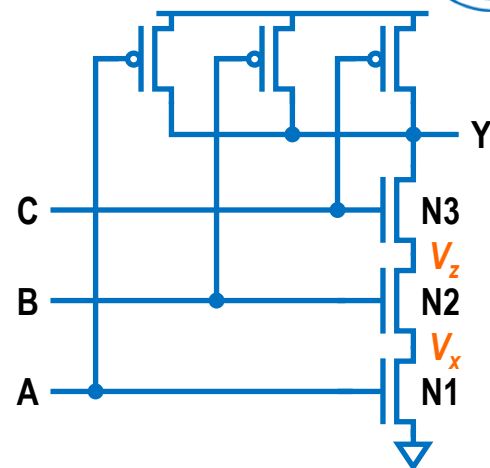
- **例：**如右图三输入与非门，参数如下：

栅氧厚度：15 Å，沟道长度：60 nm

导通NMOS管栅泄漏：6.3 nA，PMOS管栅泄漏忽略

$V_{ds}=V_{DD}$ 时，NMOS管亚阈值泄漏：5.63 nA

$|V_{ds}|=V_{DD}$ 时，PMOS管亚阈值泄漏：9.3 nA



- **解：**栅泄漏和亚阈值泄漏电流如下表所示，单位：nA

输入状态 (ABC)	I_{sub}	I_{gate}	I_{total}	V_x	V_z
000	0.4	0	0.4	堆叠效应	堆叠效应
001	0.7	0	0.7	堆叠效应	$V_{DD}-V_t$
010	0.7	1.3	2.0	中间电压	中间电压
011	3.8	0	3.8	$V_{DD}-V_t$	$V_{DD}-V_t$
100	0.7	6.3	7.0	0	堆叠效应
101	3.8	6.3	10.1	0	$V_{DD}-V_t$
110	5.6	12.7	18.3	0	0
111	28.3	19.0	47.3	0	0

■ 结泄漏

- 源/漏扩散区处于与衬底不同电位时发生
- 结泄漏与其他泄漏相比通常很小
- 高阈值电压晶体管中，BTBT和GIDL可以使结泄漏接近亚阈值泄漏水平

■ 竞争电流

- 静态CMOS逻辑电路没有竞争电流
- 有比电路、电流模式逻辑电路、模拟电路存在静态电流
- 此类电路应在休眠模式时关断

静态功耗估算举例



- **例：**考虑动态功耗估算例题中的片上系统。

电源电压：1 V；沟道长度为50 nm， $\lambda=25$ nm。芯片共有 10^9 个晶体管，其中：

逻辑管：数量50 M，平均宽度 12λ ，95%高阈值电压器件，5%低阈值电压器件；

存储管：数量950 M，平均宽度 4λ ，高阈值电压器件。

截止器件的亚阈值泄漏：低阈值电压器件100 nA/ μm ，高阈值电压器件10 nA/ μm ；

栅泄漏：5 nA/ μm ；结泄漏忽略不计。

- **求：**试估算芯片的静态功耗。

- **解：**

低阈值器件总宽度 $W_{\text{LVT}} = (50 \times 10^6) (12\lambda) (0.025 \mu\text{m}/\lambda) (0.05) = 0.75 \times 10^6 \mu\text{m}$

高阈值器件总宽度 $W_{\text{HVT}} = [(50 \times 10^6) (12\lambda) (0.95) + (950 \times 10^6) (4\lambda)] (0.025 \mu\text{m}/\lambda)$
 $= 109.25 \times 10^6 \mu\text{m}$

亚阈值泄漏电流 $I_{\text{sub}} = (W_{\text{LVT}} \times 100 \text{ nA}/\mu\text{m} + W_{\text{HVT}} \times 10 \text{ nA}/\mu\text{m})/2 = 583.75 \text{ mA}$

栅泄漏电流 $I_{\text{gate}} = [(W_{\text{LVT}} + W_{\text{HVT}}) \times 5 \text{ nA}/\mu\text{m}]/2 = 275 \text{ mA}$

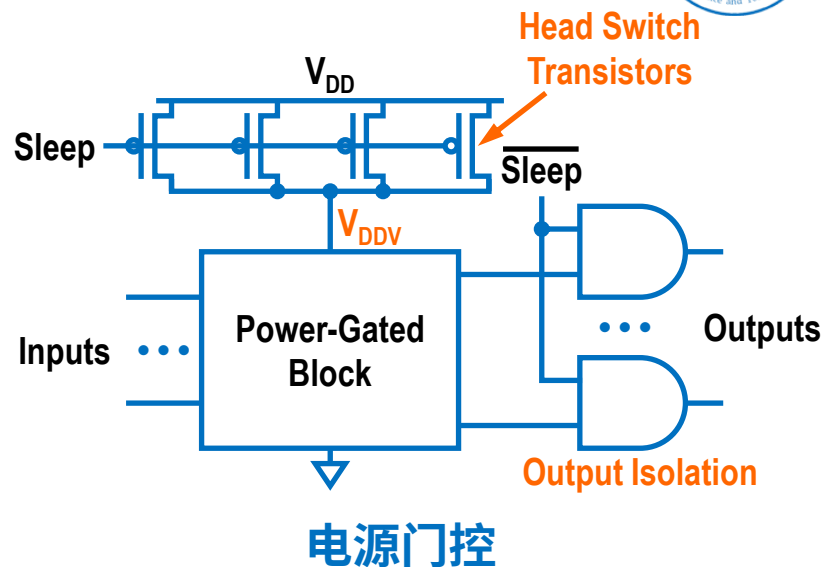
静态功耗 $P_{\text{static}} = (584 \text{ mA} + 275 \text{ mA}) (1.0 \text{ V}) = 858.75 \text{ mW}$

■ 电源门控 (Power Gating)

- 关断休眠模块的电源
- 由虚拟电源 V_{DDV} 供电
- 输出门控以免无效电平传至下游电路

■ 电源门控设计

- 模块状态的保持或恢复
 - 状态保持寄存器，使用第二电源保持状态
 - 重要寄存器内容存入存储器中，并在恢复供电时从存储器重新载入
- 电源门控粒度
 - 细粒度(Fine-Grained): 对单个逻辑门进行电源门控
 - 粗粒度(Coarse-Grained): 整个模块共享一个电源门控开关
- 开关管尺寸设计
 - 开关管上的电压降增加正常工作的延时，开关管尺寸需要足够大以减少影响
 - 宽开关管的翻转造成较大的动态功耗，只有电路休眠时间足够长才比较有效



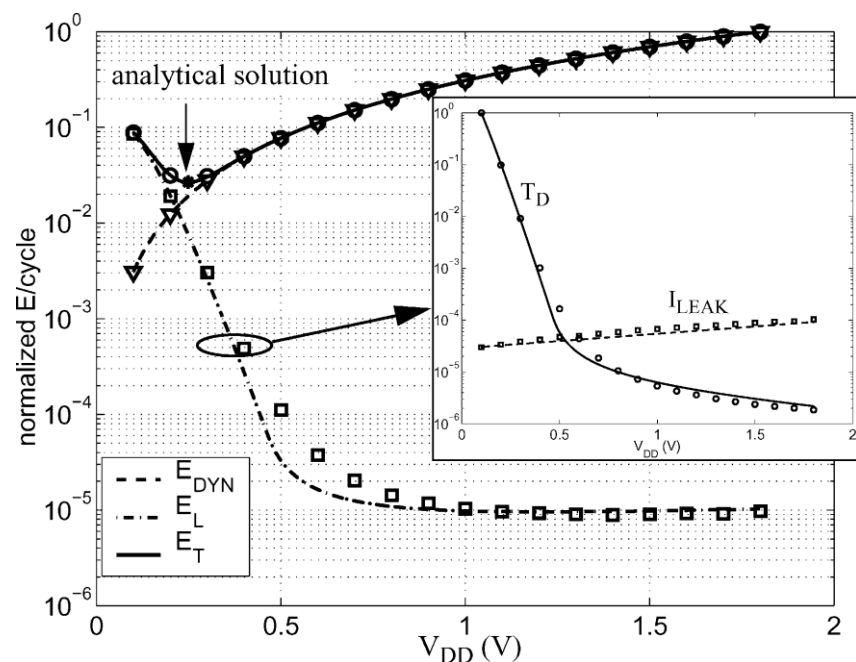
能量-延时优化

■ 功耗-延时积

- Power-Delay Product, PDP
- 一个操作的功耗与其完成时间之积，即该操作的能耗

■ 最小能耗工作点

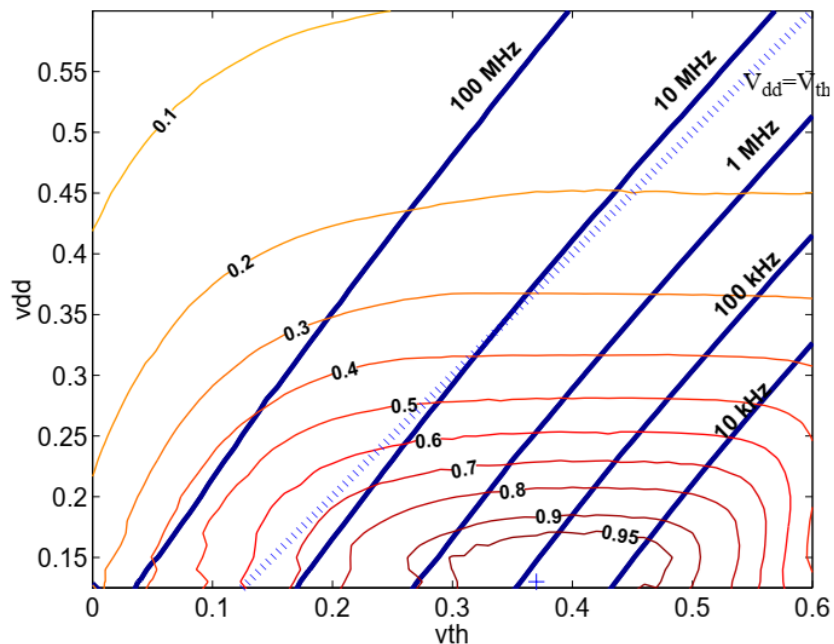
- 在不考虑延时的情况下，一个操作可能消耗的最小能量
- 发生于 $V_{DD} < V_t$ 的亚阈值工作状态
- 栅泄漏、结泄漏以及短路功耗可以忽略不计
- 总能耗为翻转能耗和泄漏能耗之和，在二者曲线交叉点处附近达到最小



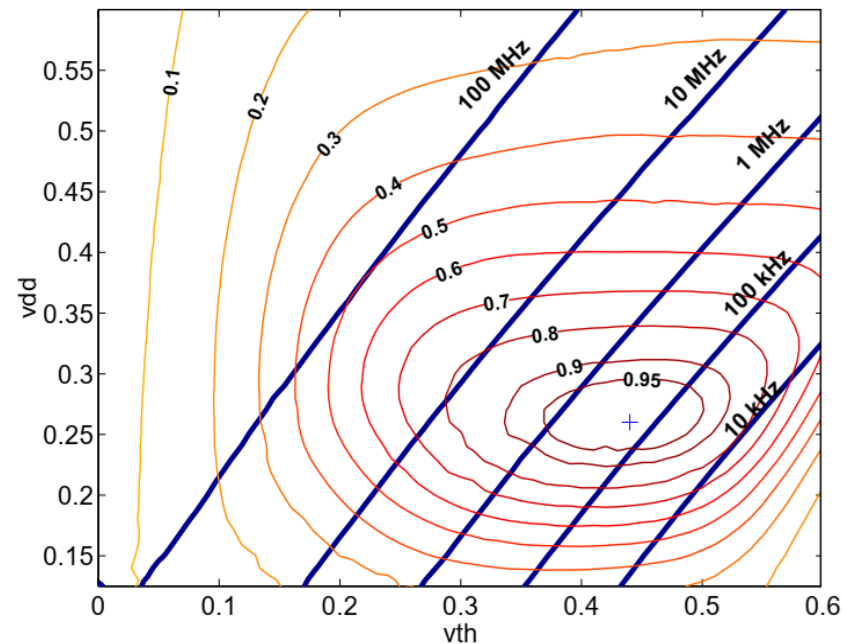
最小能耗工作点

(JSSC, 40(9), 2005, 1778–1786)

$\alpha = 1$



$\alpha = 0.1$



环形振荡器的能耗和延时等值线

等值线定义：最小能耗与能耗之比

(ISVLSI, 2002, 7-11)

■ 能耗-延时积

- Energy-Delay Product, EDP
- 均衡能耗和延时重要性的常用度量标准

■ 最小能耗-延时积

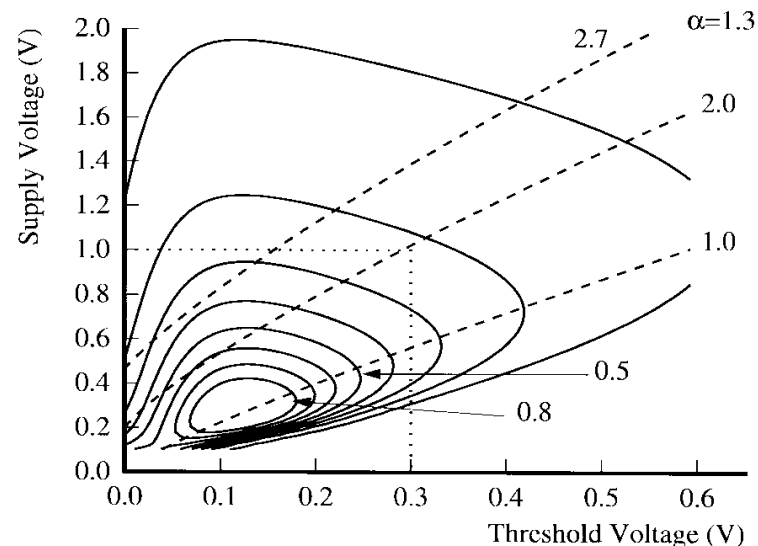
- 若忽略泄漏，由 α 幂律模型可得

$$\text{EDP} = k \frac{C_{\text{eff}}^2 V_{\text{DD}}^3}{(V_{\text{DD}} - V_t)^\alpha}$$

- 使EDP取最小值的电源电压为

$$V_{\text{DD-opt}} = \frac{3}{3-\alpha} V_t, \quad 1 \leq \alpha \leq 2$$

- 若考虑泄漏，可通过EDP等值线进行分析



能耗-延时积的等值线

(JSSC, 32(8), 1997, 1210-1216)

低功耗体系结构

■ 处理器

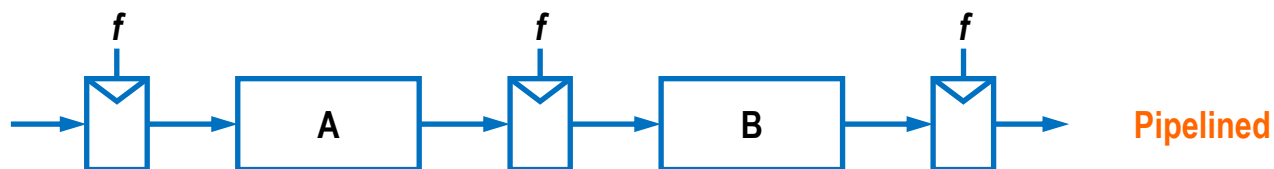
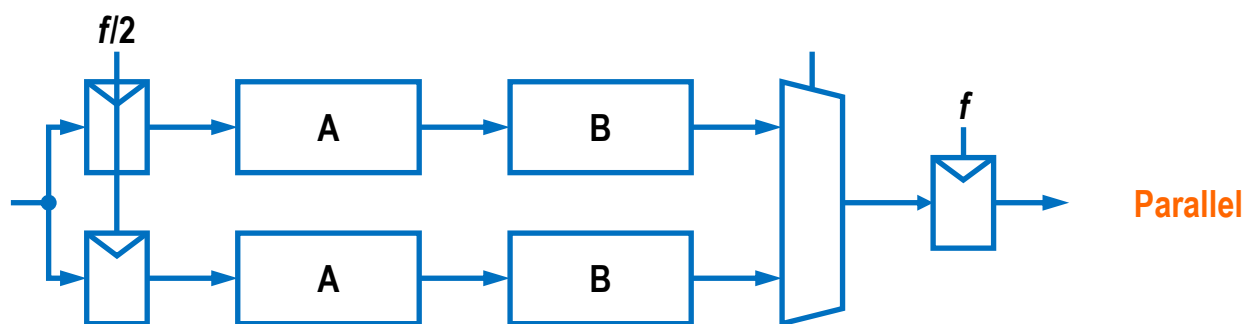
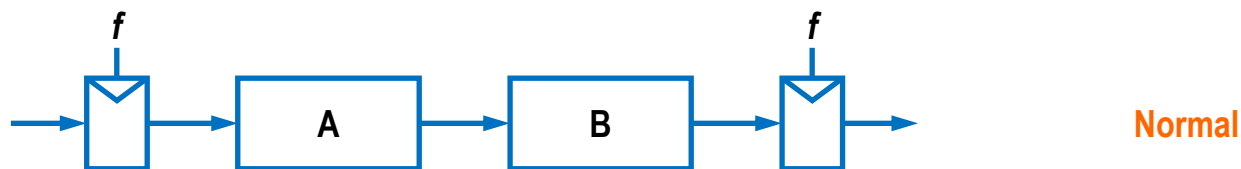
- 处理器性能随晶体管数量的平方根增长
- 采用较多数量的简单内核去处理任务和数据级的并行性
- 较小的内核具有较短的连线和较快的存储器访问速度

■ 存储器





























- 存储器具有比逻辑低得多的功耗密度
- 存储器活动因子很小并且具有规整性，可以简化对泄漏的控制
- 优先考虑采用较大的存储器而非较快的处理器进行任务加速

■ 专用功能单元

- 专用功能单元提供的能量效率通常比通用处理器高一个数量级
- 用于高计算强度应用的加速器使处理器从这些任务中解脱出来



并行性(Parallelism)与流水线(Pipeline)

	C0 HFM	C0 LFM	C1/C2	C4	C6
Core voltage					
Core clock			OFF	OFF	OFF
PLL				OFF	OFF
L1 caches			 flushed	 flushed	 off
L2 caches				 Partial flush	 off
Wakeup time	active	active	 <1 μ s	 <30 μ s	 <100 μ s
Power					

Intel Atom处理器的电源管理模式

(JSSC, 44(1), 2009, 73–82)

本章结束