模拟与数字电路

Analog and Digital Circuits

课程主页 扫一扫

第 十四讲： **数字功耗与应用**

Lecture 14: **Digital Power and Application**

主　　讲： 陈 迟 晓

Instructor : Chixiao Chen

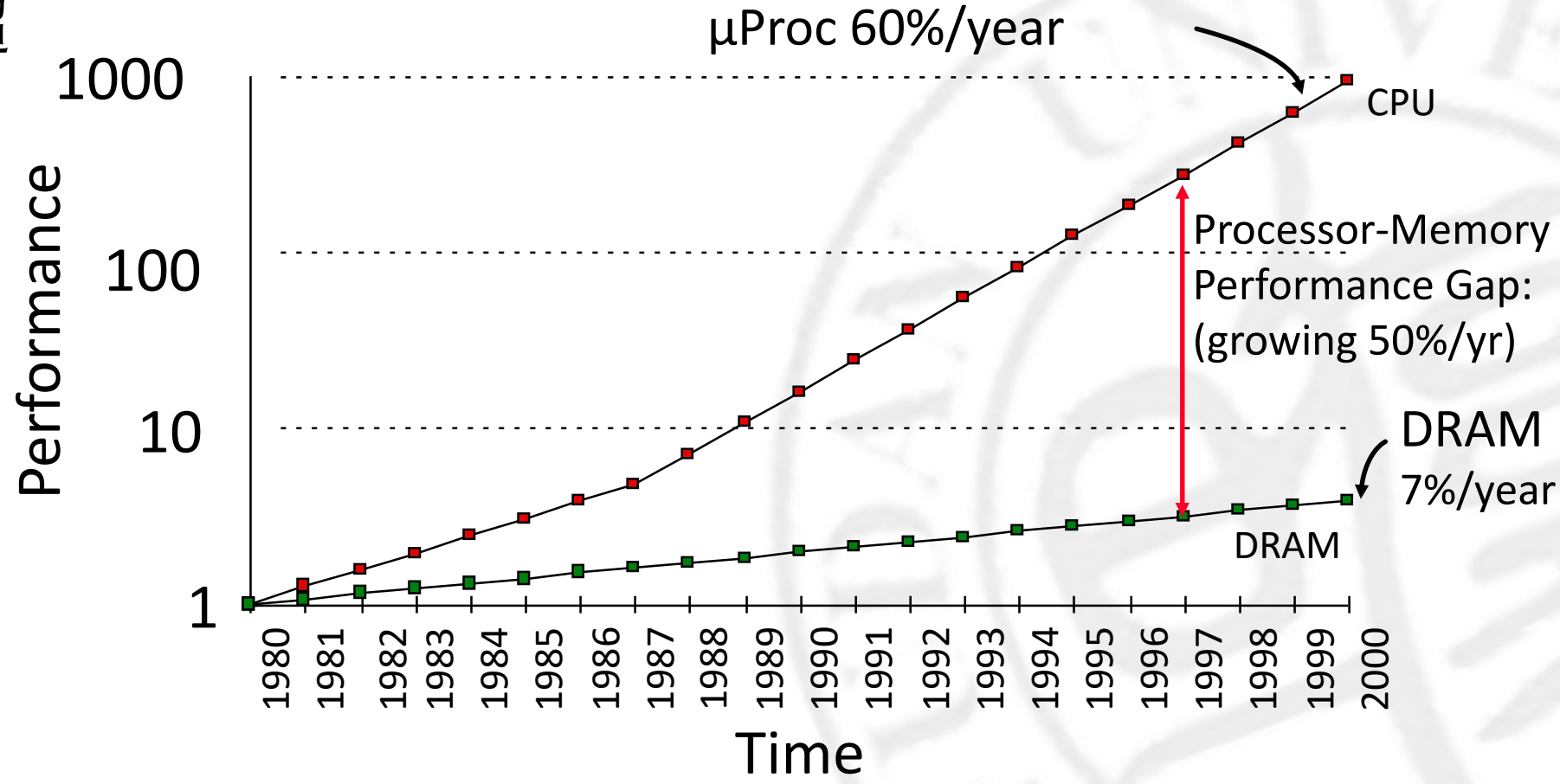
提纲

- 复习
  - 静态存储器和动态存储器的共同点和差别分别是什么？

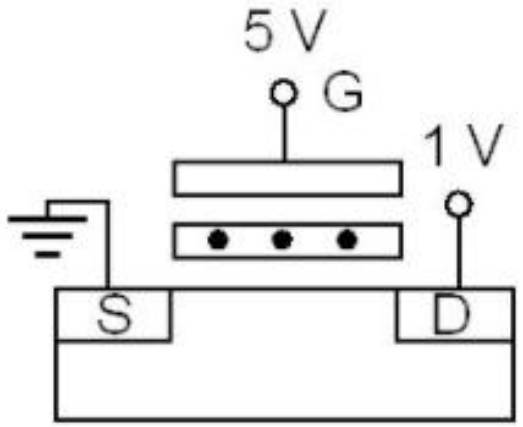
  - 非易失存储器
  - CMOS电路的功耗
  - 数字电路与处理器

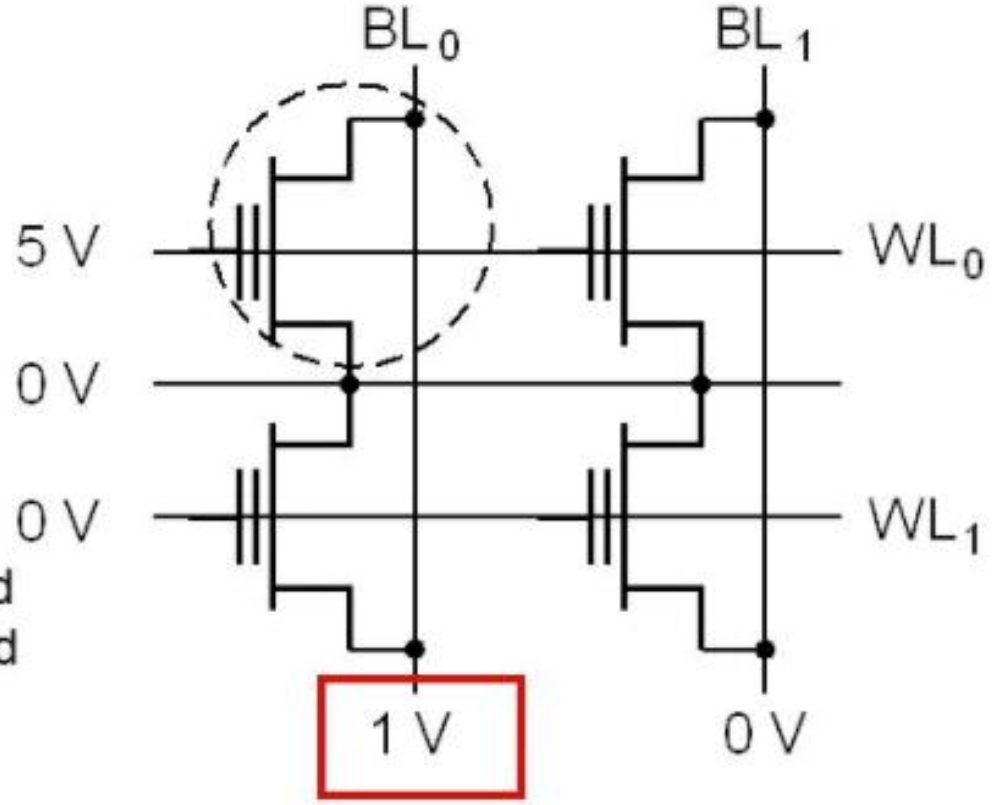
处理器-存储间的差距(延时)

- 3GHz 4-issue 超标量处理器
- 100ns DRAM
- 1200 指令



NOR Flash – 读操作



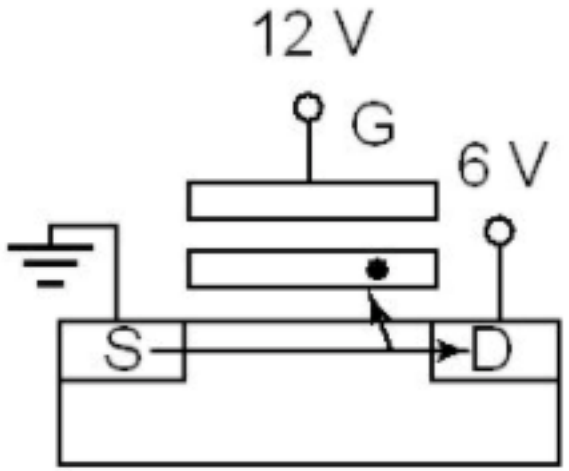


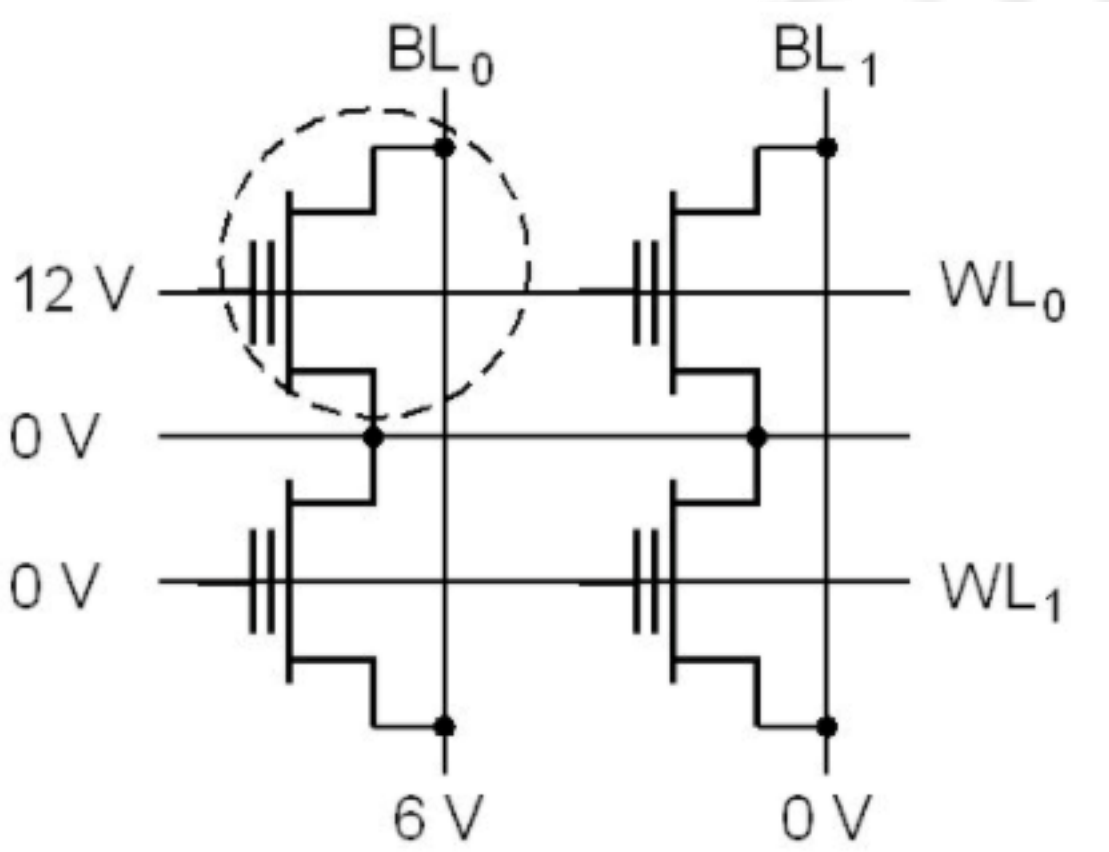
In read operation the programmed transistor stores 1 as it is switched off permanently

NOR Flash memories have

- **Fast random read time**
- **Slow erasure and programming time**
- **Need precise control of thresholds**

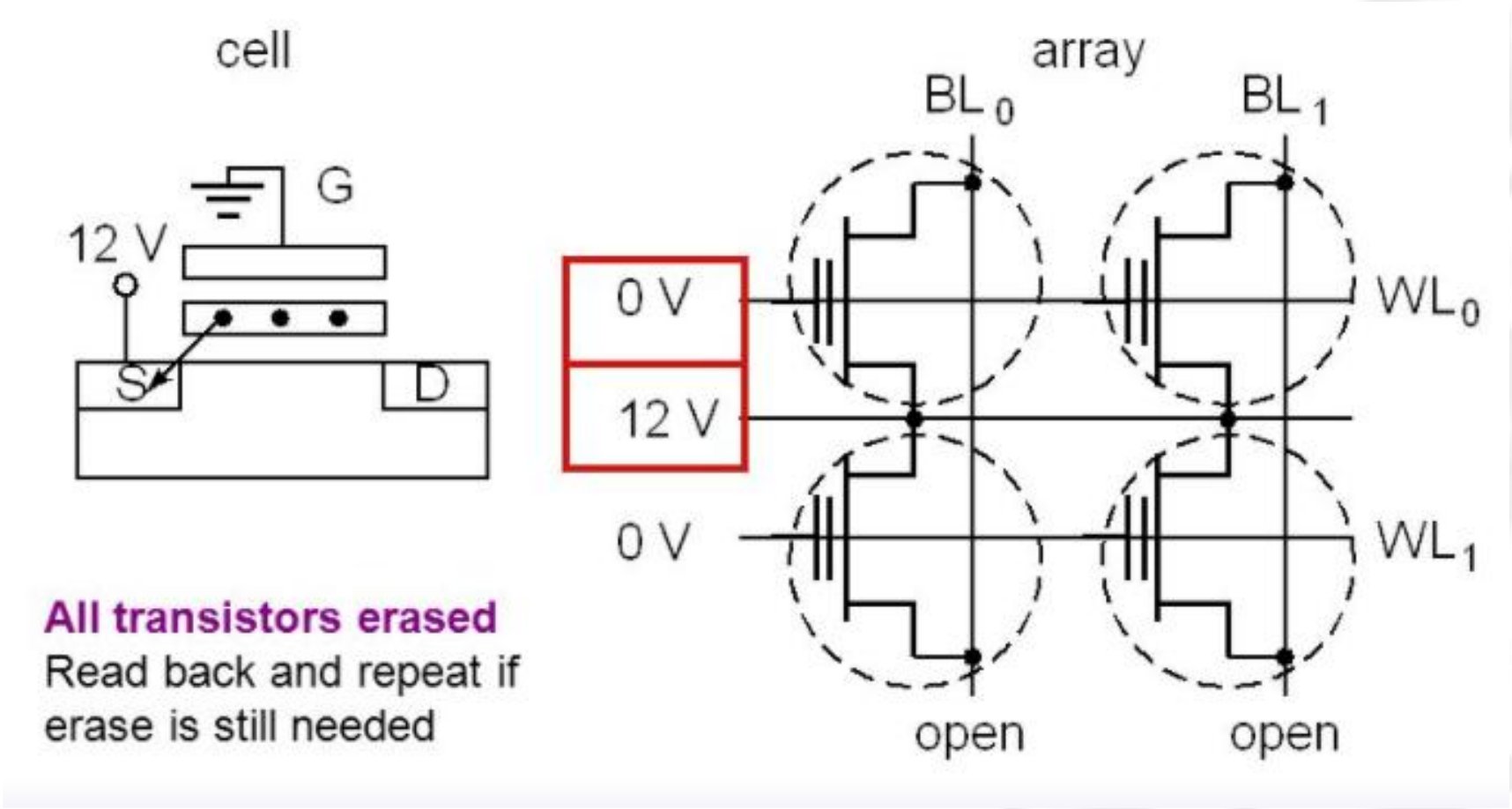
NOR Flash – 写操作





读写操作的电源电压不同

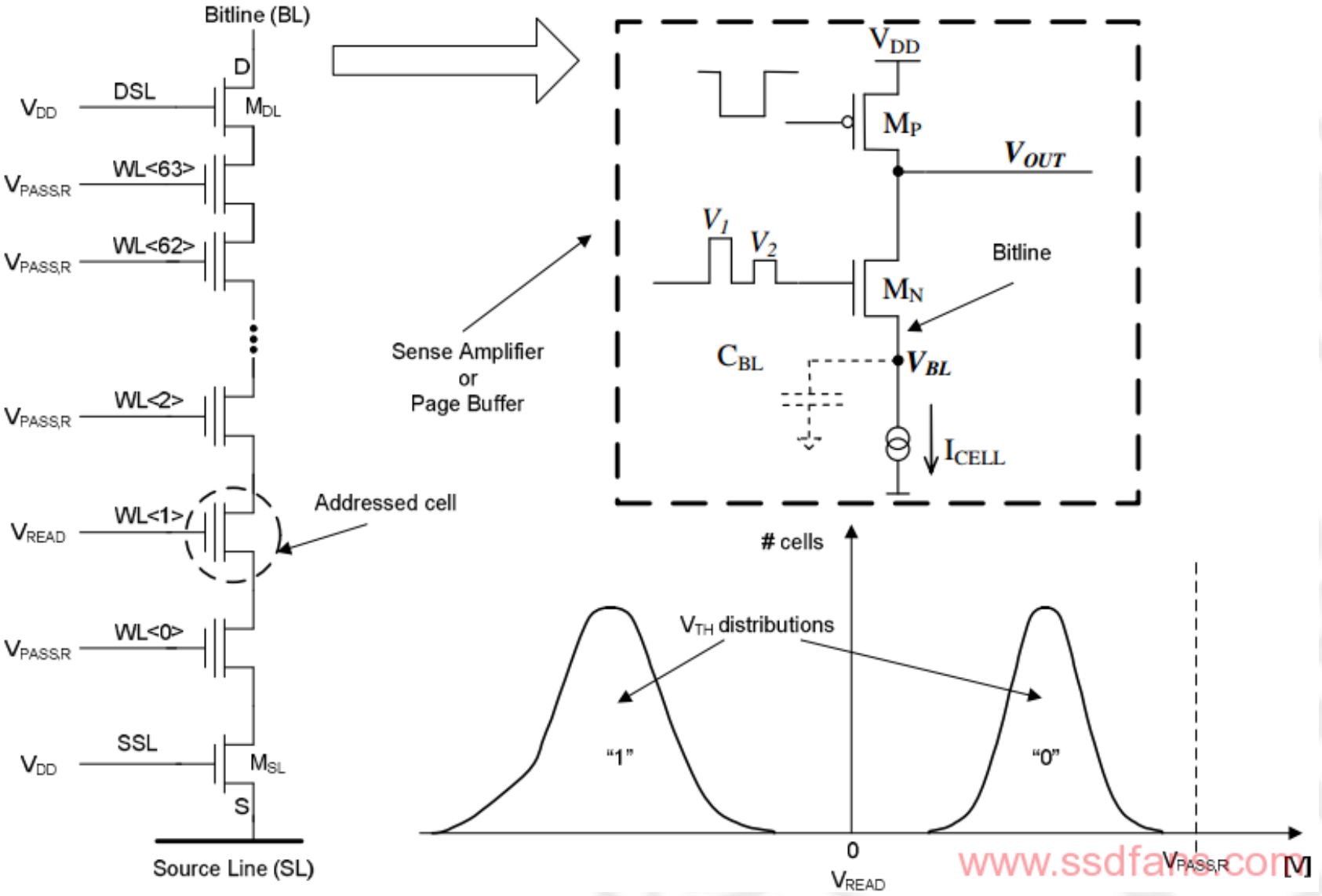
NOR Flash – 擦除操作



格式化——关注Source line

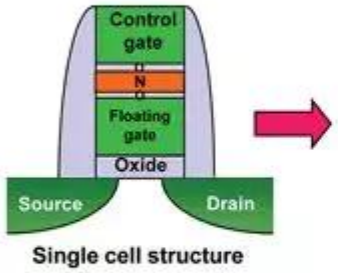
NAND Flash

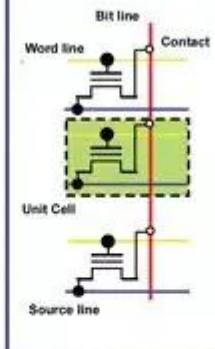
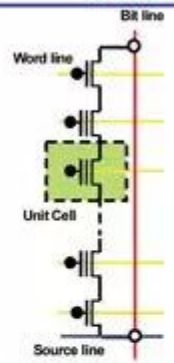
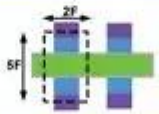
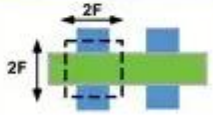
- 当我们读一个cell时它的gate施加电压Vread(0V)，其他cells的gate端加Vpass,R电压(通常4-7V)以便无论cells的Vth值大小都可以保证其他cells完全导通开启。



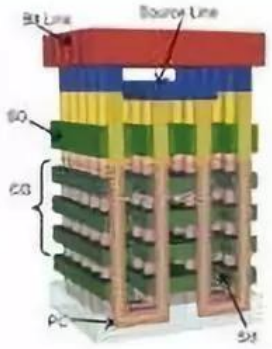
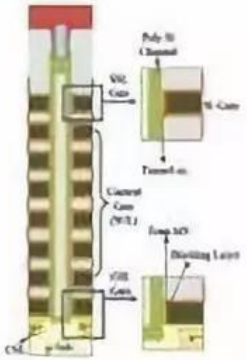
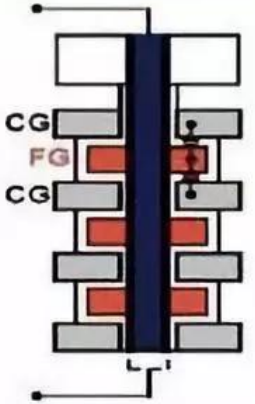
Flash SSD

- NAND Vs. NOR



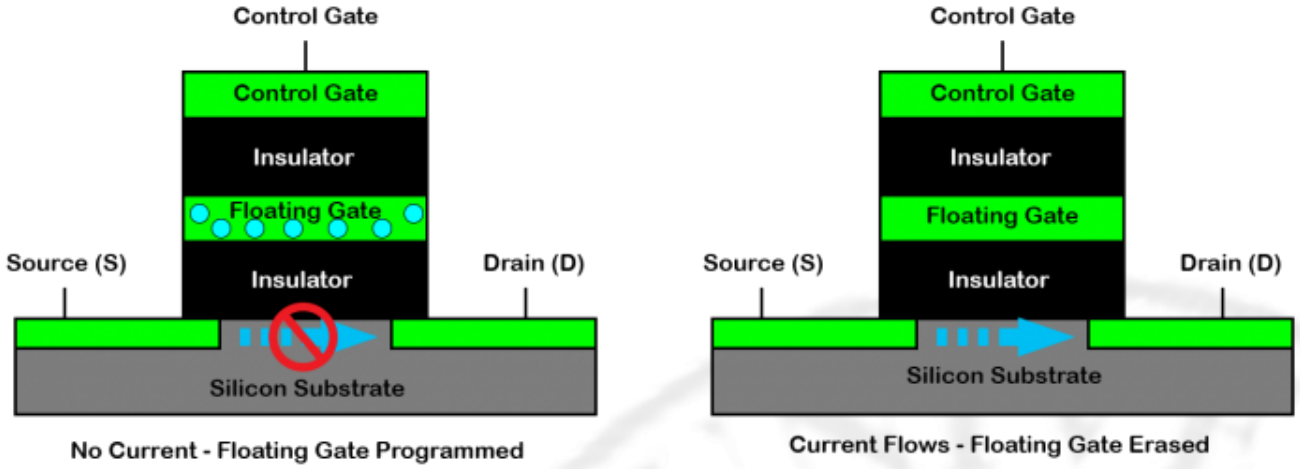
	NOR	NAND
Cell Array		
Layout		
Cross section	 $10F^2$	 $4F^2$

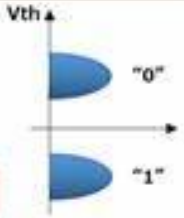
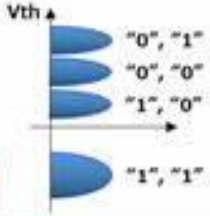
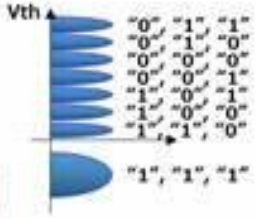
- 3D Flash

	p-BiCS (Toshiba)	TCAT (Samsung)	3D FG (Hynix)
Structure	 Tanaka. H, VLSIT 2007	 J. Jang, VLSIT 2009	 S. Whang, IEDM 2010
Key Features	- P+ SONOS Cell	- TANOS Cell	- Floating Gate
Key Issue	- Large Cell Size - Reliability	- Large Cell Size - SL Resistance	- Process of bit separation - Disturbance

Flash SSD

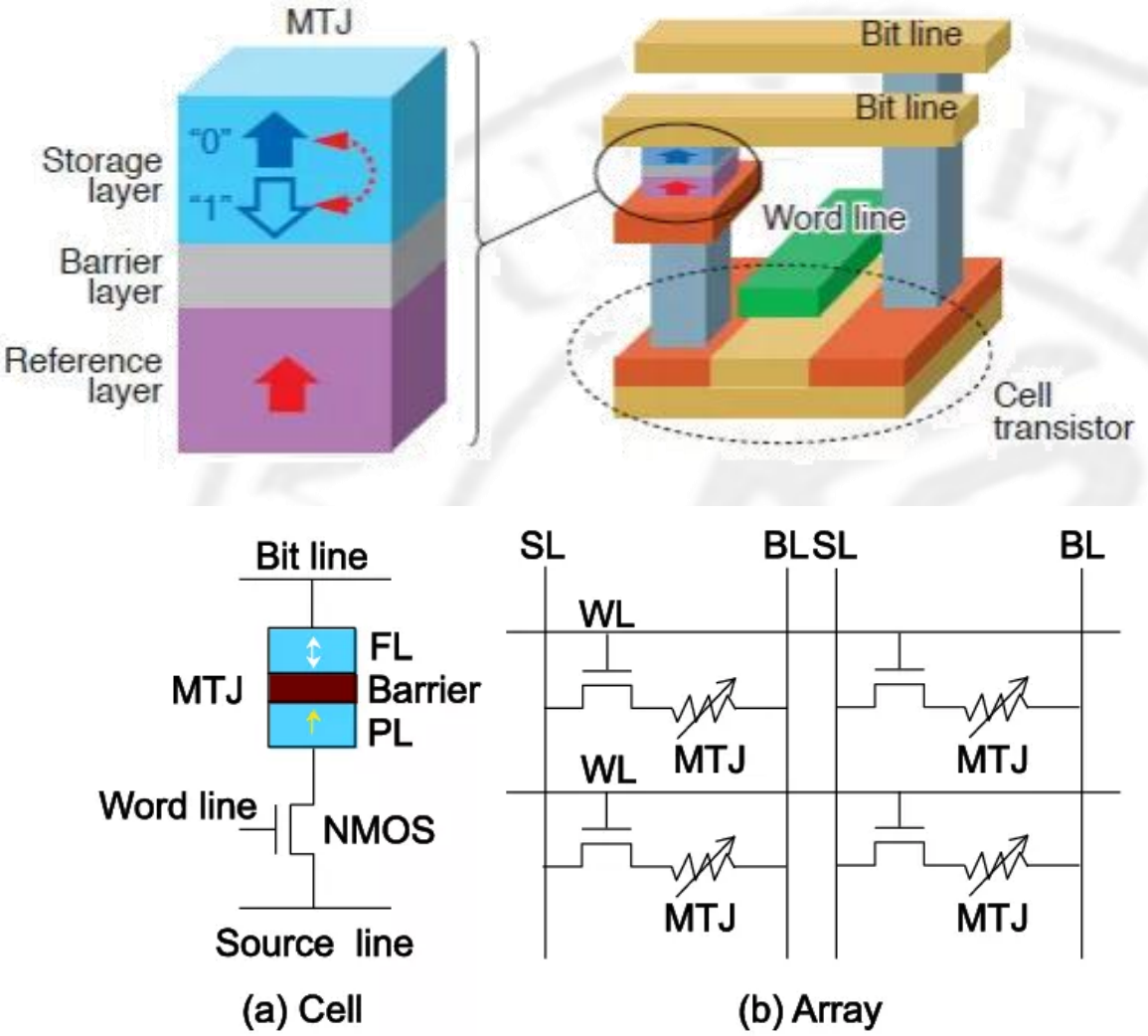
- 在浮栅中充电
- 单值单元 vs. 多值单元



Device Type	Stored Information / Memory Cell	State Count	Vth Distribution of the memory cell
SLC (Single Level Cell)	**1** bit / cell	**2**	
1 bit per cell → Reliable, Higher cost			
MLC (Multi Level Cell)	**2** bits / cell	**4**	
2 bits share same cell → doubled Capacity , less reliable			
TLC (Triple Level Cell)	**3** bits / cell	**8**	
3 bits share same cell → Higher Capacity , poor reliable			

新原理存储器 – MRAM/ReRAM

- 易失与非易失存储
- 与DRAM一样快/密集，与SRAM一样兼容CMOS，与闪存一样非易失?
- STT-MRAM(Spin Transfer Torque Magnetic RAM)自旋转移扭矩随机存取存储器
- 使用忆阻器的ReRAM



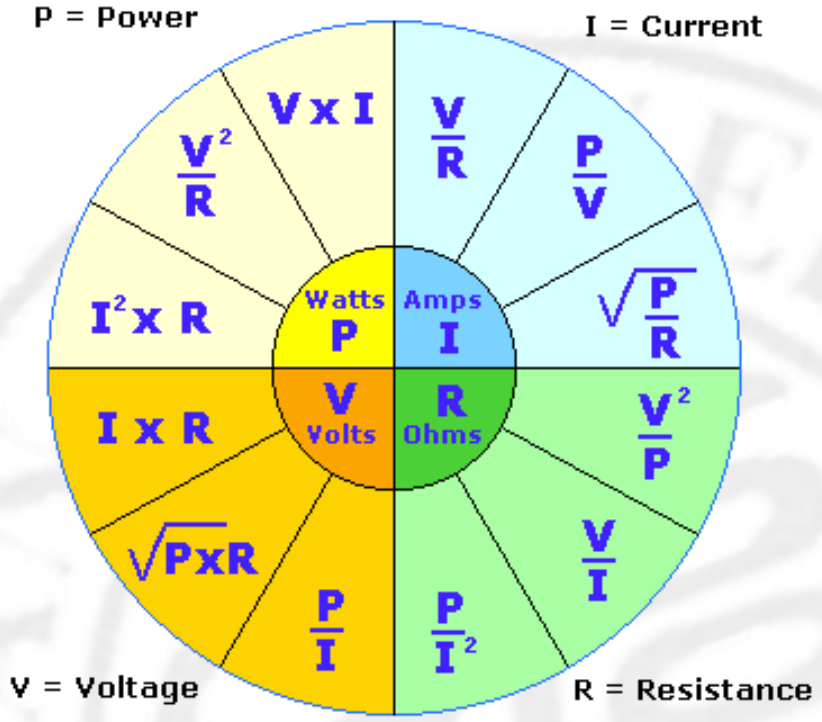
Energy (能耗) vs. Power (功耗)

- Energy is the ability to do work
  - 单位：Joule

$$P = \frac{dW}{dt}$$

- Power is rate of expending energy
  - 单位：Watt

- Energy Efficiency: energy per operation
  - 单位： Op per seconde per Watt OPS/W



CMOS数字电路的功耗推导

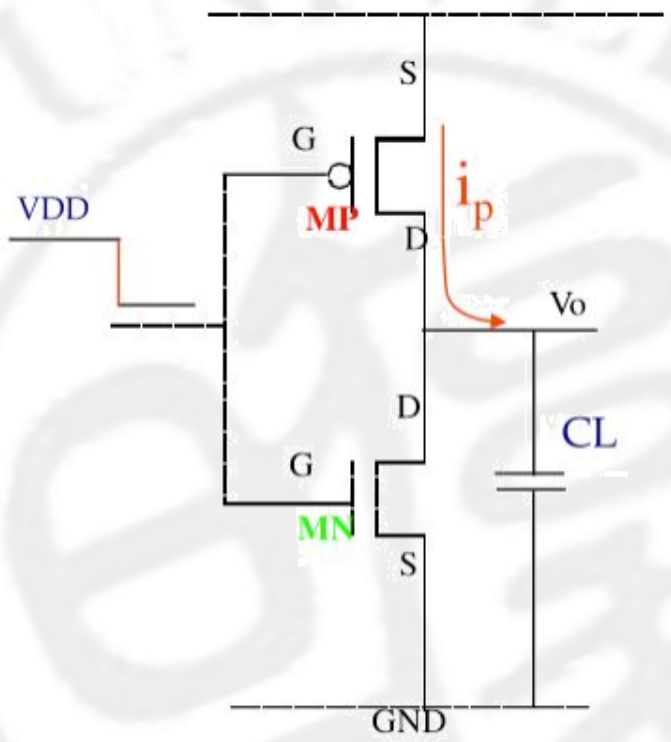
- 以反相器为例子
- 每次从1→0
- 开关功耗

$$P_{dp} = \frac{1}{tp}\int_0^{t1} ip(t)(VDD - Vo)dt$$

$$ip(t) = C_L \frac{dVo}{dt}$$

$$Pdp = \frac{C_L}{tp}\int_0^{VDD} (VDD - Vo)dVo$$

$$Pdp = \frac{C_L}{2tp}(VDD)^2$$



CMOS数字电路芯片功耗

动态功耗（开关）

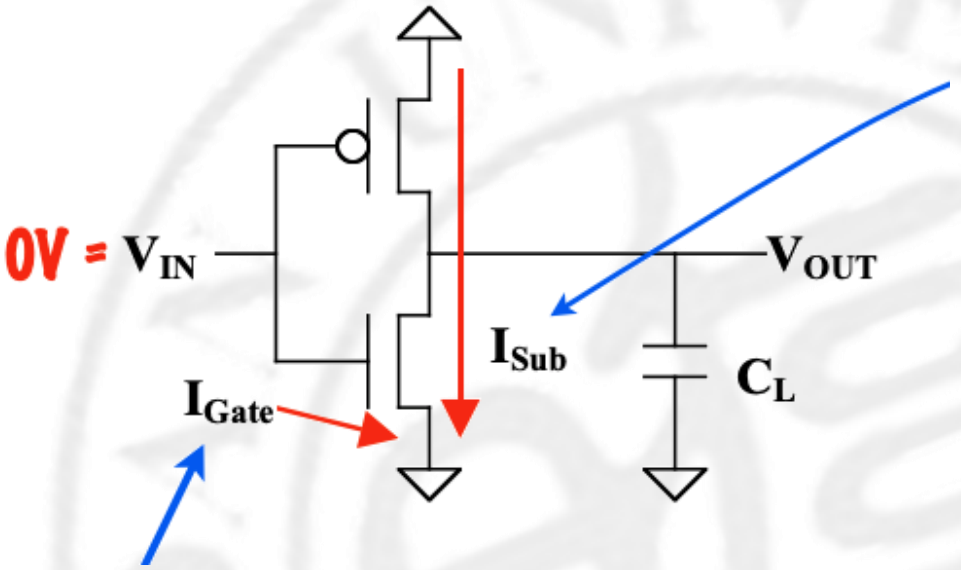
$$P_{sw} = 1/2\ \alpha\ C\ V_{dd}^{2}\ F$$

"activity factor", average percentage of capacitance switching per cycle (~ number of nodes to switch)

Total chip capacitance to be switched

Clock Frequency

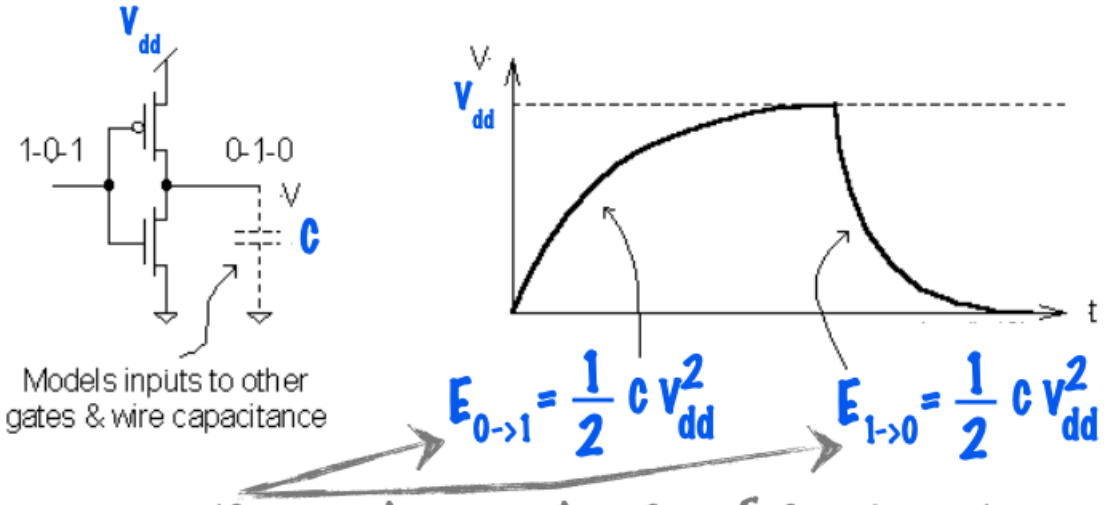
静态功耗（漏电）



理想条件下，开关关断，电阻无限大

实际情况：纳米尺寸晶体管，电阻有限大
每个晶体管的漏电在nA，若超大规模集成
那么漏电也不可忽略

如何高能效：摩尔定律



Strong result: Independent of technology.

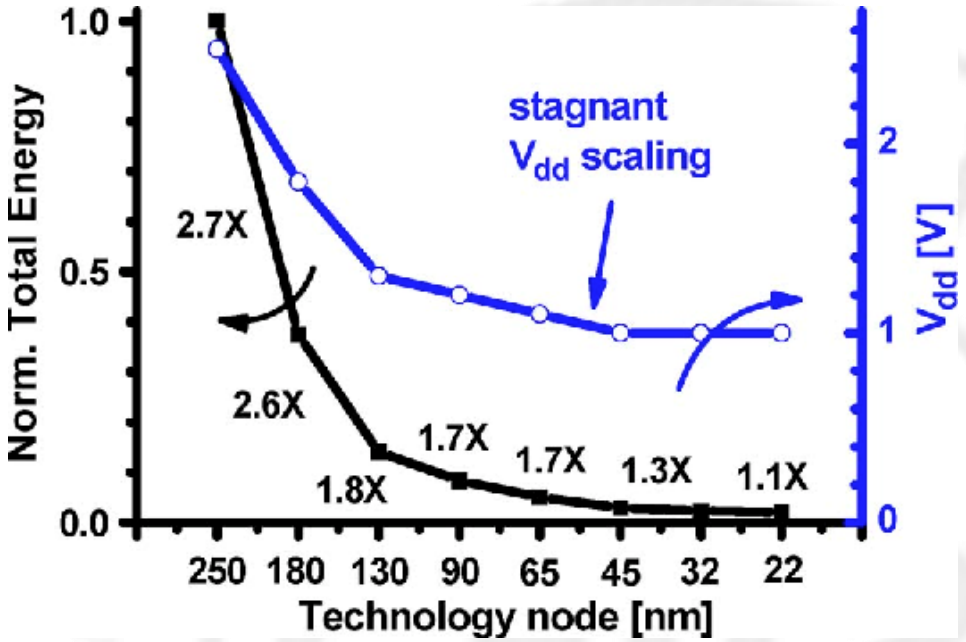
How can we limit switching energy?

(1) Reduce # of clock transitions. But we have work to do ...

(2) Reduce Vdd. But lowering Vdd limits the clock speed ...

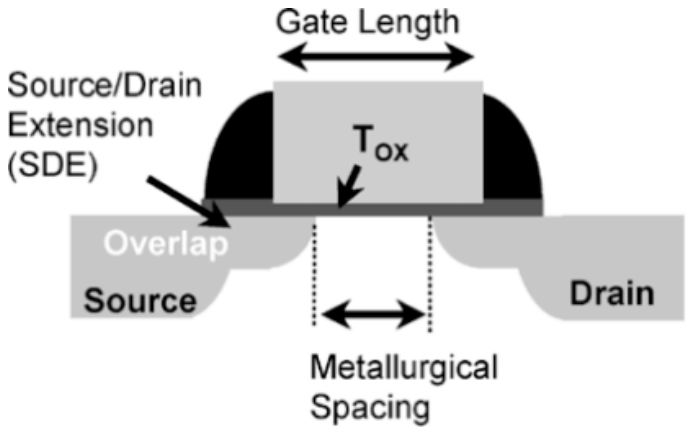
(3) Fewer circuits. But more transistors can do more work.

(4) Reduce C per node. One reason why we scale processes.



器件微缩（微纳器件）可以同时降低Vdd与负载电容，还能提高速度

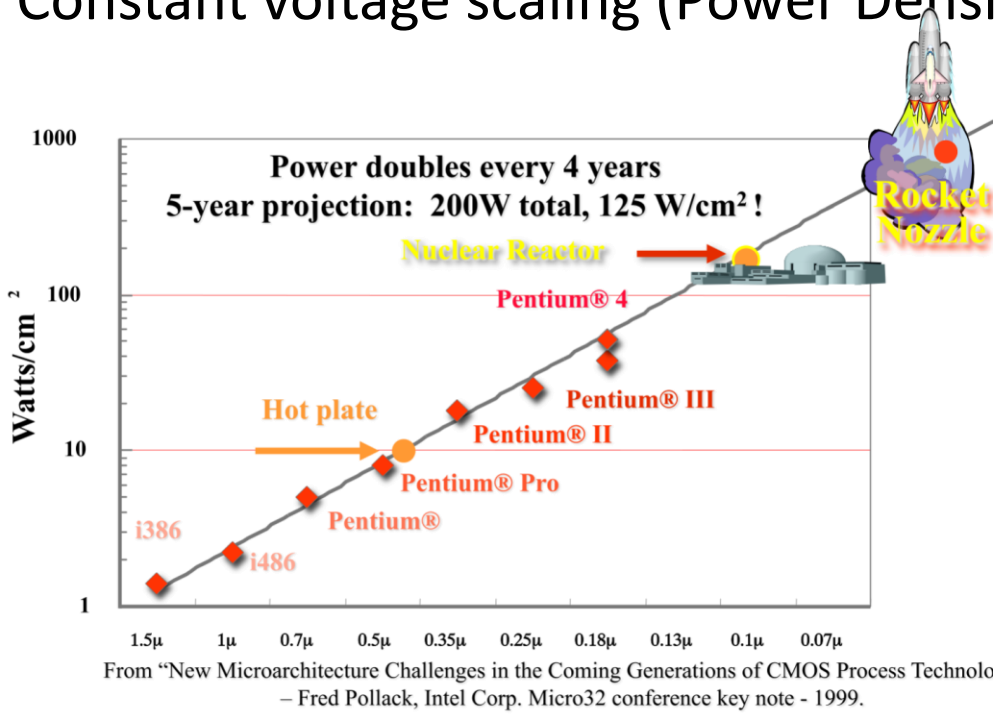
Full Scaling vs. Dennard Scaling



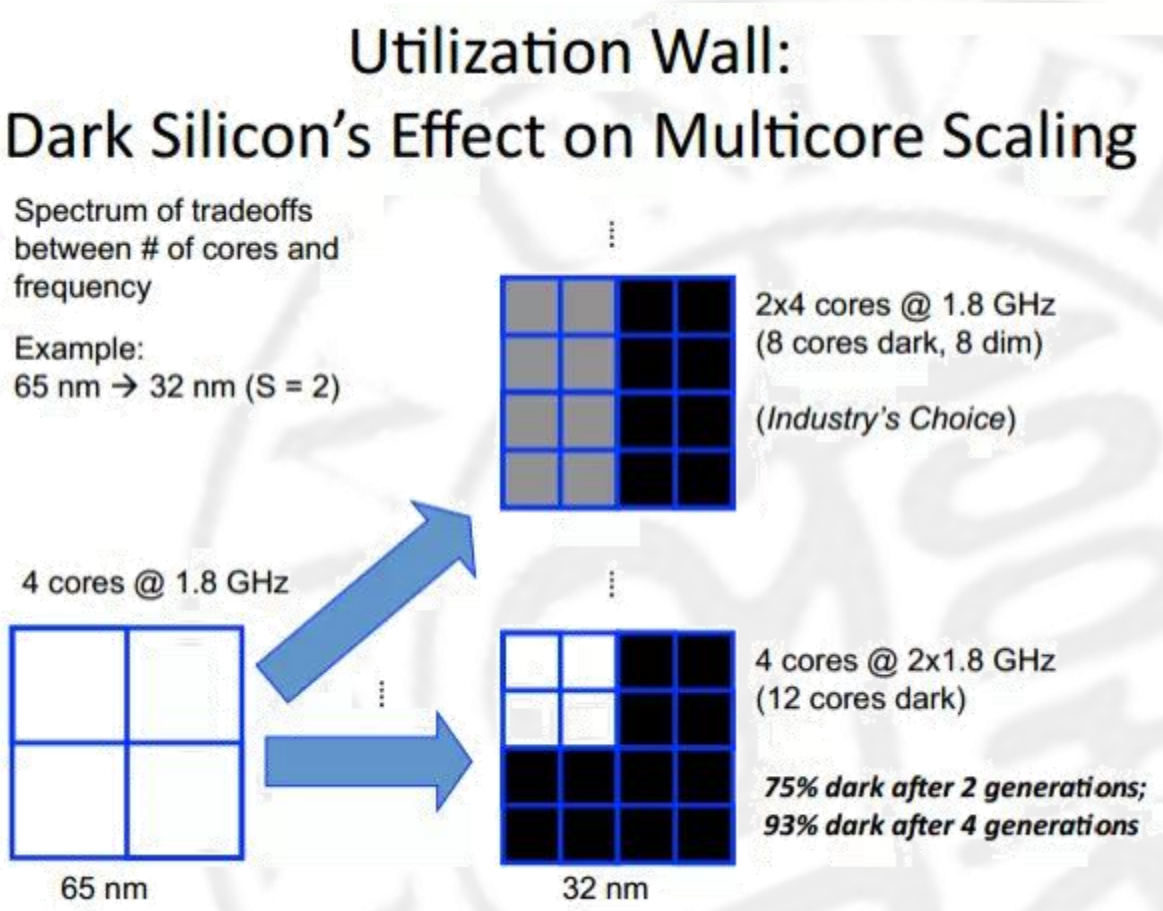
Parameters	Before Scaling	After scaling	
		Full scaling	Constant Voltage scaling
Channel length	L	$L' = L/s$	$L' = L/s$
Channel width	W	$W' = W/s$	$W' = W/s$
Gate oxide thickness	tox	$tox' = tox/s$	$tox' = tox/s$
Junction depth	Xj	$Xj' = Xj/s$	$Xj' = Xj/s$
Power supply voltage	VDD	$VDD' = VDD/s$	$VDD' = VDD$
Threshold voltage	VT0	$VT0' = VT0/s$	$VT0' = VT0$
Doping densities	N_A, N_D	$N_A', N_D' = sN_A , sN_D$	$N_A', N_D' = s^2 N_A , s^2 N_D$
Oxide capacitance	Cox	$Cox' = Cox/s$	$Cox' = Cox/s$
Drain current	I_D	$I_D' = I_D/s$	$I_D' = s.I_D$
Power dissipation	P_D	$P_D' = P_D/s^2$	$P_D' = s. P_D$
Power density	P_D/Area	P_D/Area' = P_D/Area	P_D/Area' = s^3. P_D/Area

Post Moore's Law Era —— Dark Silicon

- Theorical deviation of Moore's Law
  - Constant field Scaling (Dennard)
  - Constant voltage scaling (Power Density)



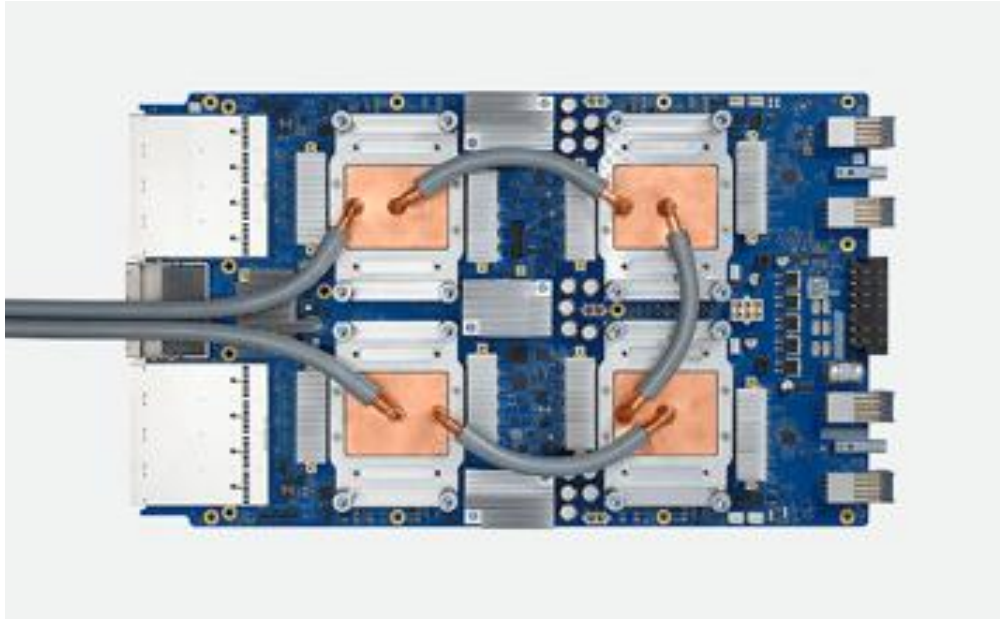
From "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technolo
– Fred Pollack, Intel Corp. Micro32 conference key note - 1999.

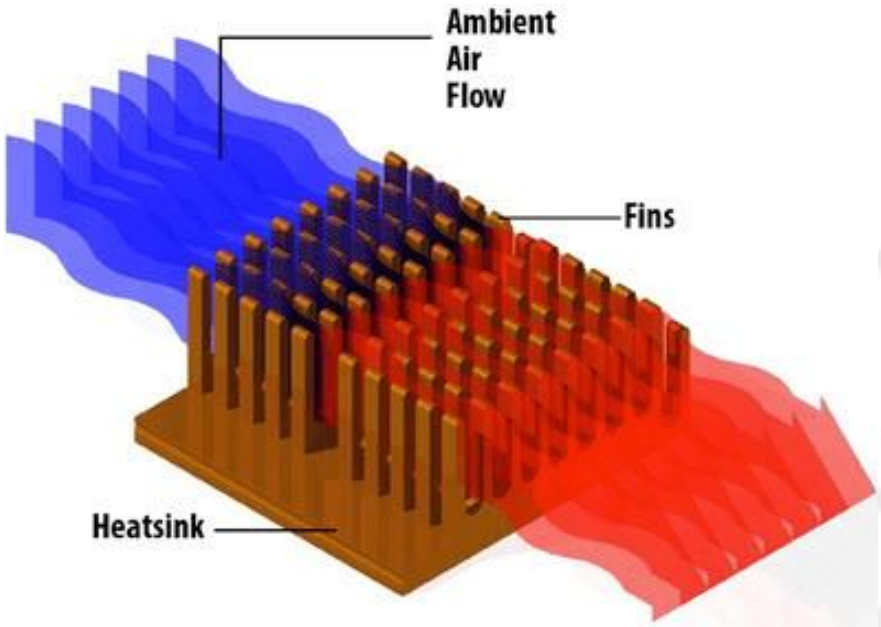


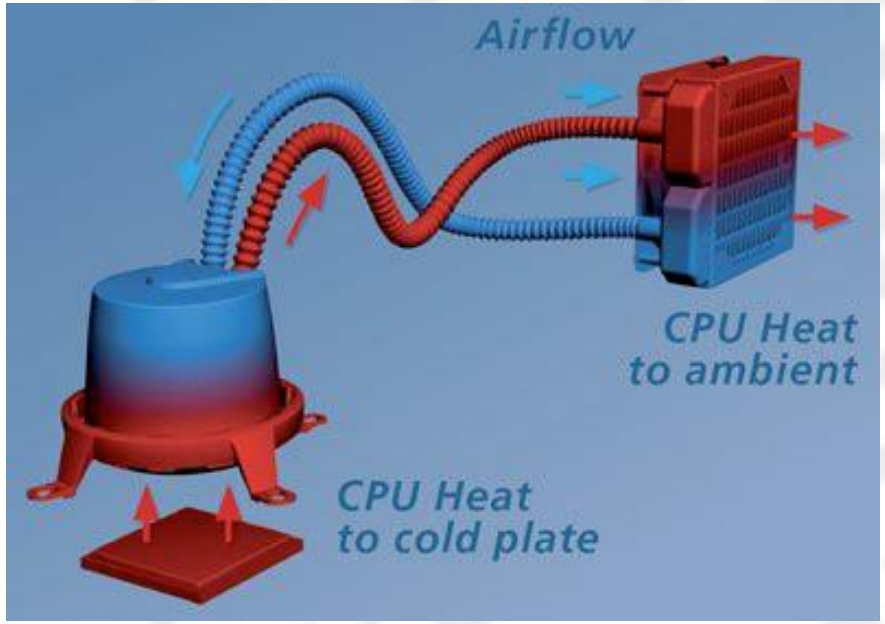
Cores cannot work simultaneously with power limits,

散热

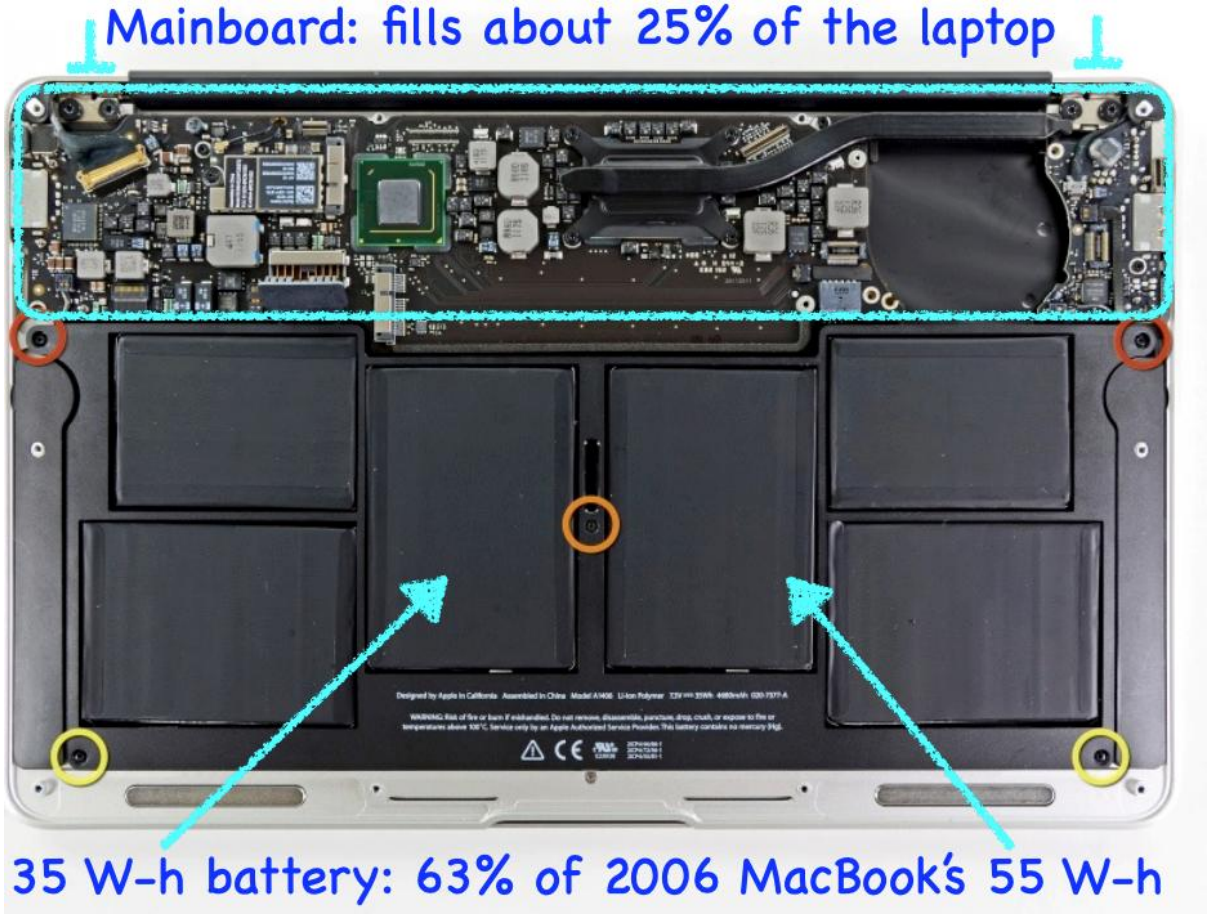
- 功耗与热量是伴生的，但是晶体管无法工作在高温状态
- 典型散热方式：
  - 风冷\液冷







数字电路的供电方式



大规模数据中心
散热与供电功耗
接近甚至超过计
算功耗

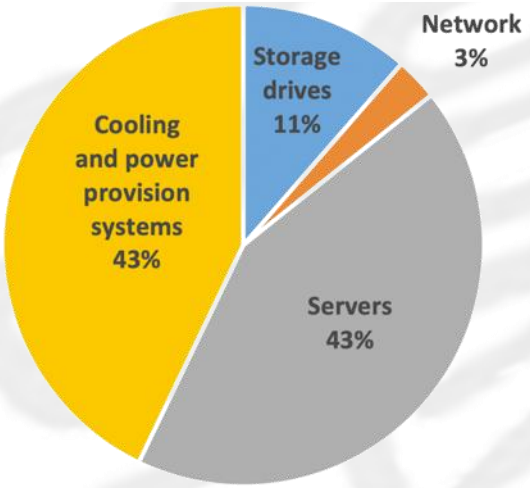
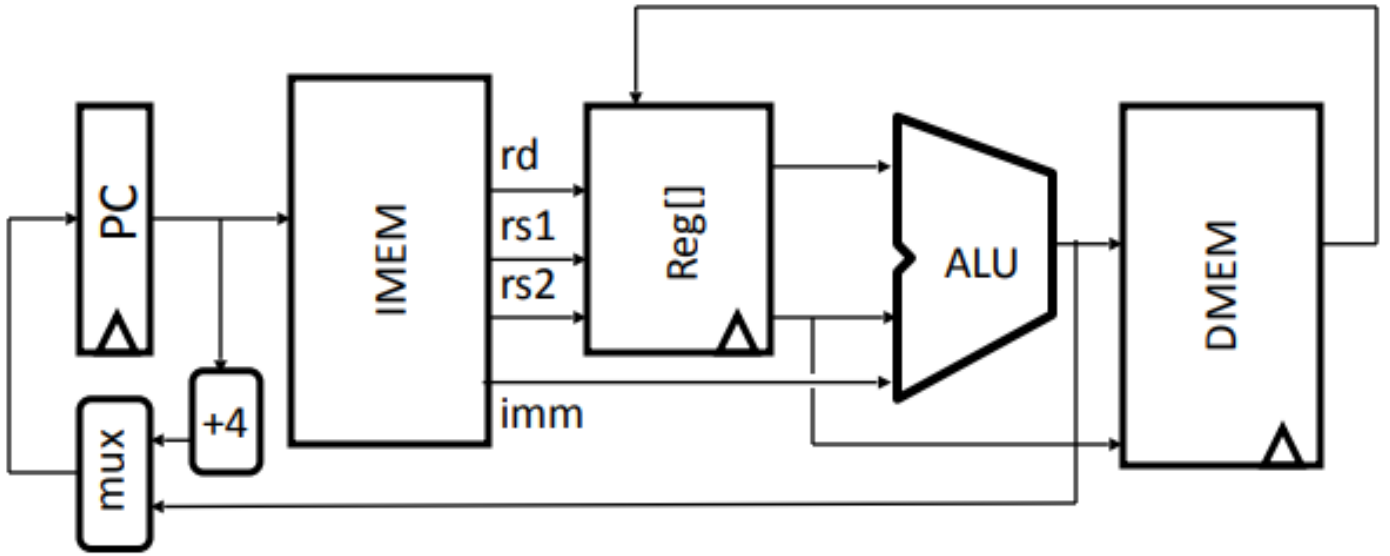
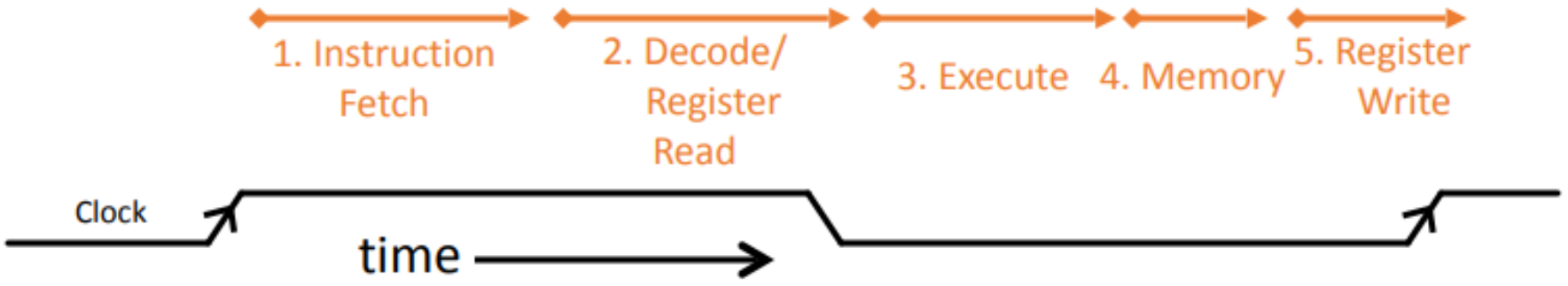


Figure 1. Fraction of U.S. data center electricity use in 2014, by end use. Source: Shehabi 2016.

基于数字电路的处理器设计

处理器是用于按照规定执行顺序计算机指令集的数字电路

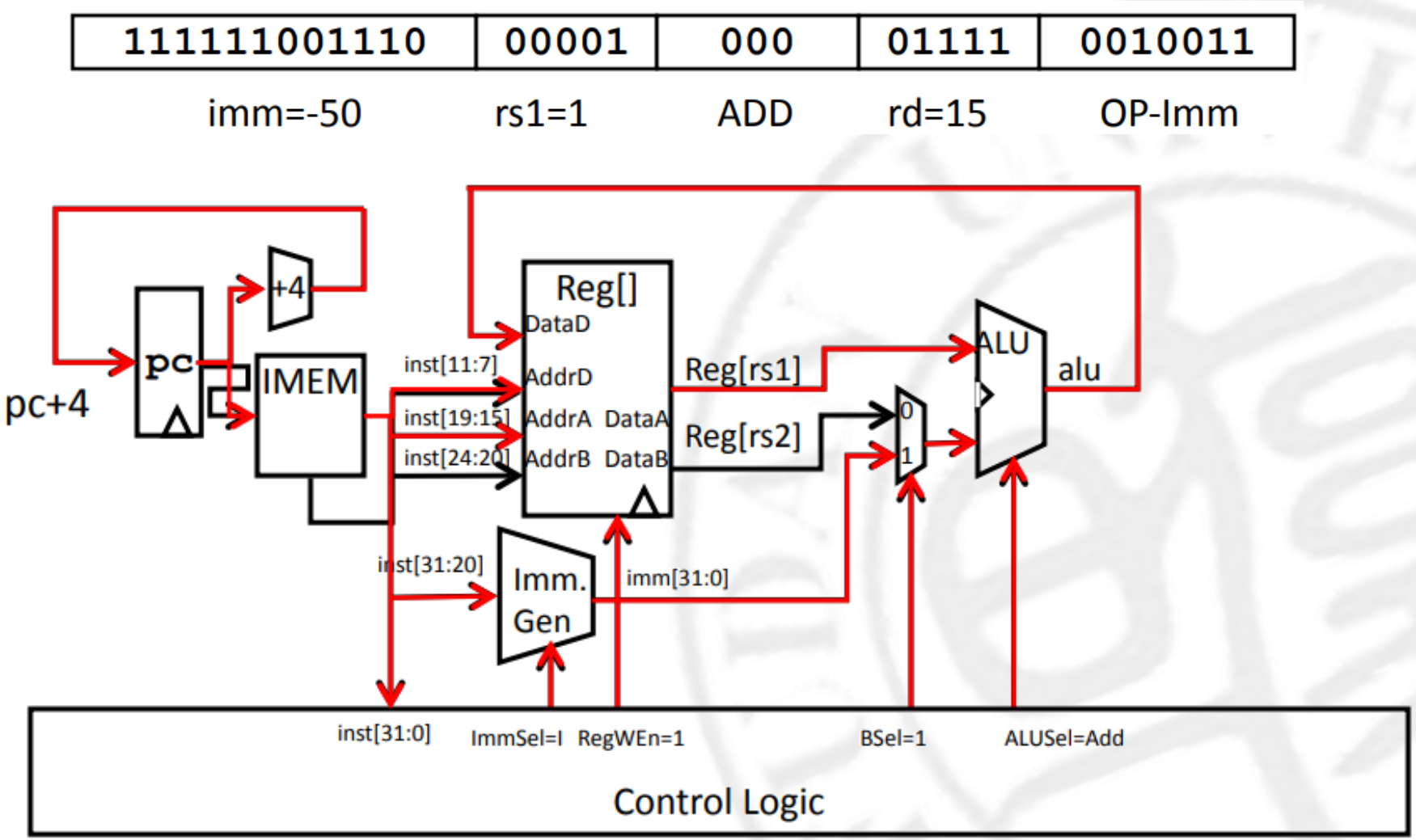




完成一条加法指令addi 的电路实现

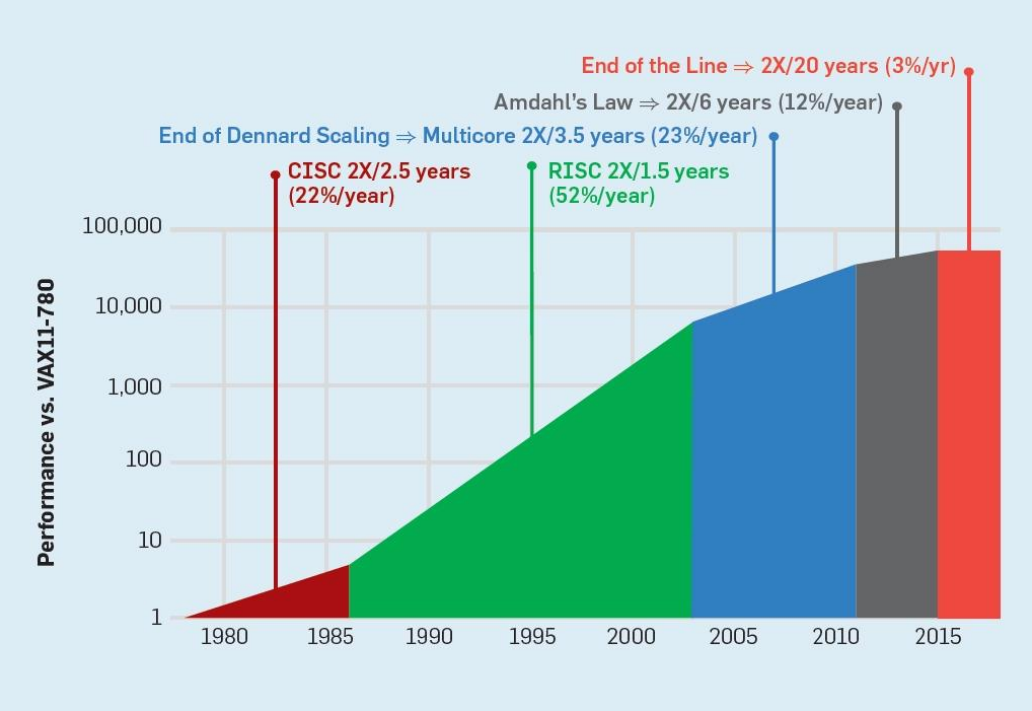
汇编代码:

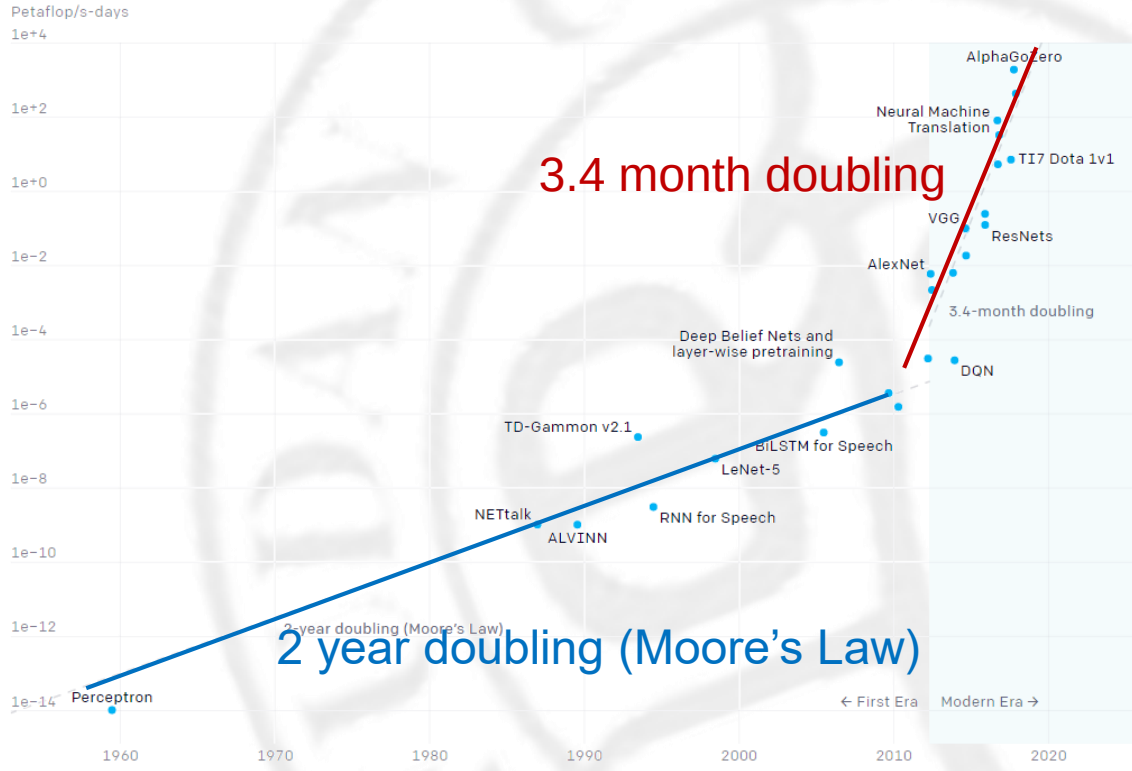
addi x15,x1,-50



数字电路遇到的全新挑战

General purpose processors (CPUs) can **NOT** afford the recent high performance computing brought by fast growing AI algorithms.





Architecture solution for Dark Silicon

- Leverage dark silicon to fight the utilization wall
  - Power is more expensive than area now
  - Specialized logic can achieve 10-1000x better energy efficiency
  - General purpose designs waste 90% power, use **specialized** ones instead !