The Abdus Salam International Centre for Theoretical Physics

2499-7

International Training Workshop on FPGA Design for Scientific Instrumentation and Computing

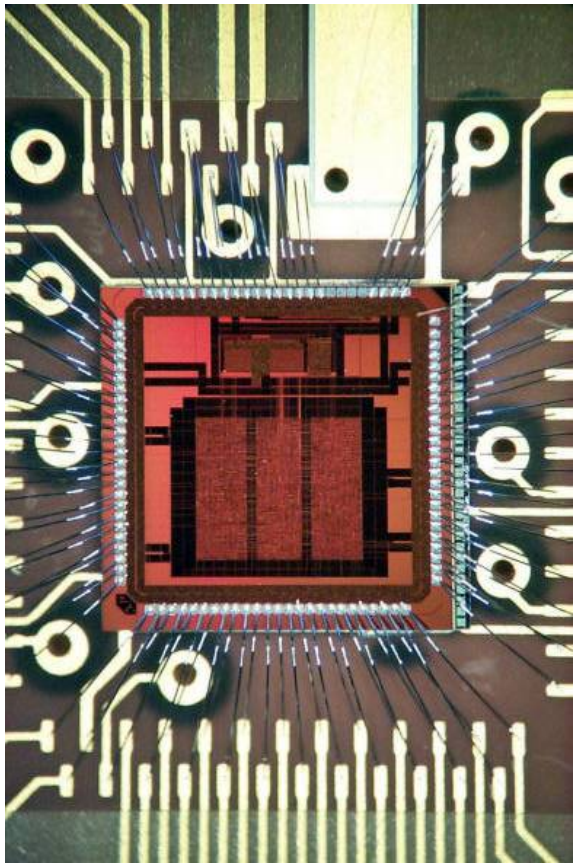
11 - 22 November 2013

Introduction to VLSI Digital Design
Scaling

Sandro BONACINI
CERN, Geneva
Switzerland

Outline

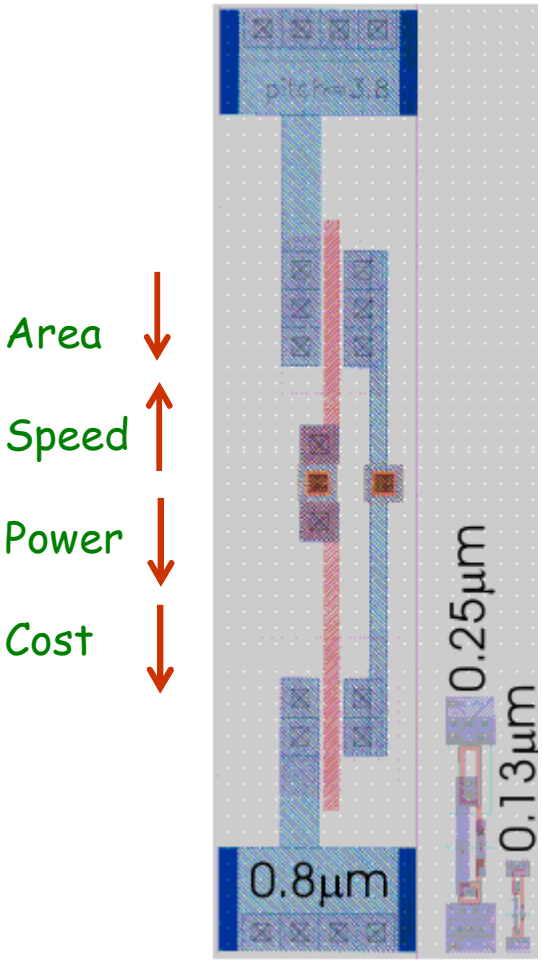
- Introduction
- Transistors
- The CMOS inverter
- Technology
- Scaling
  - Scaling objectives
  - Scaling variables
  - Scaling consequences
- Gates
- Sequential circuits
- Storage elements



Many slides are a courtesy of Paulo Moreira

Technology scaling

- Technology scaling has a threefold objective:
  - Increase the transistor density
  - Reduce the gate delay
  - Reduce the power consumption
- At present, between two technology generations, the objectives are:
  - Doubling of the transistor density;
  - Reduction of the gate delay by 30% (43% increase in frequency);
  - Reduction of the power by 50% (at 43% increase in frequency);



Technology scaling

- How is scaling achieved?
  - All the device dimensions (lateral and vertical) are reduced by $1/\alpha$
  - Concentration densities are increased by α to make the junctions depletion region smaller by $1/\alpha$
  - Device voltages reduced by $1/\alpha$ (for constant field)
  - Typically $1/\alpha = 0.7$ (30% reduction in the dimensions)

Technology scaling

- The **scaling variables** are:
  - Supply voltage: $V_{dd} \rightarrow V_{dd} / \alpha$
  - Gate length: $L \rightarrow L / \alpha$
  - Gate width: $W \rightarrow W / \alpha$
  - Gate-oxide thickness: $t_{ox} \rightarrow t_{ox} / \alpha$
  - Junction depth: $X_j \rightarrow X_j / \alpha$
  - Substrate doping: $N_A \rightarrow N_A \times \alpha$

  This is called **constant field** scaling because the electric field across the gate-oxide does not change when the technology is scaled

  If the power supply voltage is maintained constant the scaling is called **constant voltage**. In this case, the electric field across the gate-oxide increases as the technology is scaled down.

Due to gate-oxide breakdown, below 0.8μm only "constant field" scaling is used.

Scaling consequences

Some consequences of 30% scaling in the constant field regime (α = 1.43, $1/\alpha$ = 0.7):

- Device/die area:

$$W \times L \rightarrow (1/\alpha)^2 = 0.49$$

  - "Historically", microprocessor die size grows about 25% per technology generation! This is a result of added functionality.

- Transistor density:

$$(\text{unit area}) / (W \times L) \rightarrow \alpha^2 = 2.04$$

  - In practice, memory density has been scaling as expected.

Scaling consequences

- Gate capacitance:

$$W \times L / t_{ox} \rightarrow 1/\alpha = 0.7$$

- Drain current:

$$(W/L) \times (V^2/t_{ox}) \rightarrow 1/\alpha = 0.7$$

- Gate delay:

$$(C \times V) / I \rightarrow 1/\alpha = 0.7$$
$$\text{Frequency} \rightarrow \alpha = 1.43$$

  - In practice, microprocessor frequency has doubled every technology generation (2 to 3 years)! This faster increase rate is due to super-pipelined architectures ("less gates per clock cycle")

Scaling consequences

- Power:

$$C \times V^2 \times f \rightarrow (1/\alpha)^2 = 0.49$$

- Power density:

$$1/t_{ox} \times V^2 \times f \rightarrow 1$$

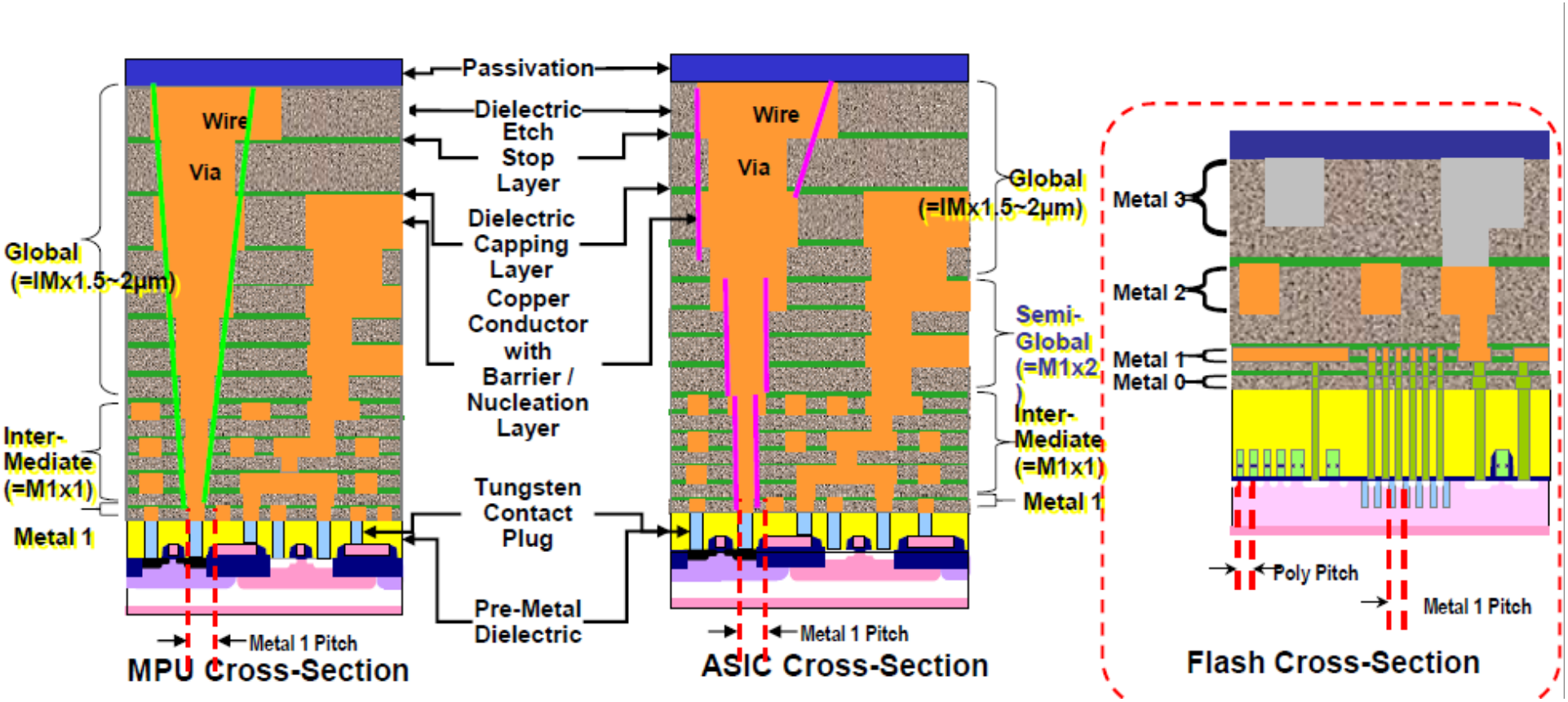
  - In practice, the power density has been increasing faster than foreseen by the simple scaling theory. This is due to the faster than foreseen increase in frequency

Interconnects scaling

- Interconnects scaling:
  - Higher densities are only possible if the interconnects also scale.
  - Reduced width → increased resistance
  - Denser interconnects → higher capacitance
  - To account for increased parasitics and integration complexity **more interconnection layers** are added:
    - thinner and tighter layers → local interconnections
    - thicker and sparser layers → global interconnections and power

IC Cross Sections

Source: ITRS



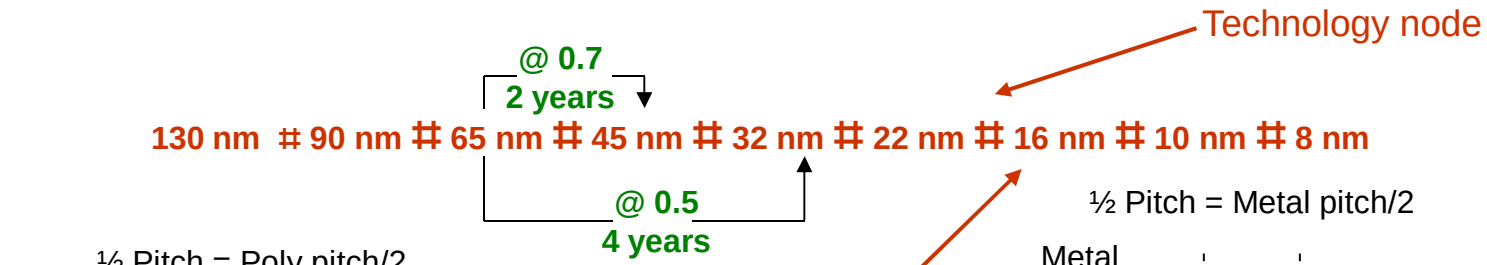
Scaling table

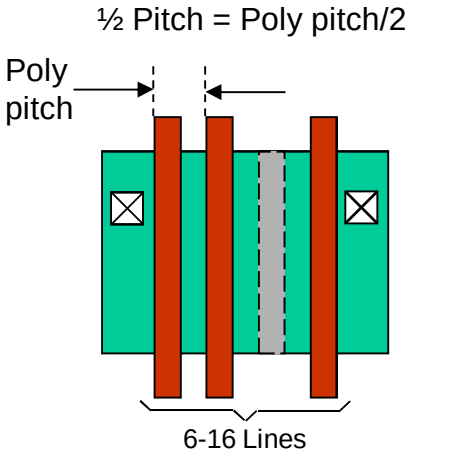
Parameter	Constant Field	Constant Voltage	
Supply voltage (V_{dd})	$1/\alpha$	1	Scaling Variables
Length (L)	$1/\alpha$	$1/\alpha$	
Width (W)	$1/\alpha$	$1/\alpha$	
Gate-oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$	
Junction depth (X_j)	$1/\alpha$	$1/\alpha$	
Substrate doping (N_A)	α	α	
Electric field across gate oxide (E)	1	α	Device Repercussion
Depletion layer thickness	$1/\alpha$	$1/\alpha$	
Gate area (Die area)	$1/\alpha^2$	$1/\alpha^2$	
Gate capacitance (load) (C)	$1/\alpha$	$1/\alpha$	
Drain-current (I_{dss})	$1/\alpha$	α	
Transconductance (g_m)	1	α	
Gate delay	$1/\alpha$	$1/\alpha^2$	Circuit Repercussion
Current density	α	α^3	
DC & Dynamic power dissipation	$1/\alpha^2$	α	
Power density	1	α^3	
Power-Delay product	$1/\alpha^3$	$1/\alpha$	

2013 and beyond …

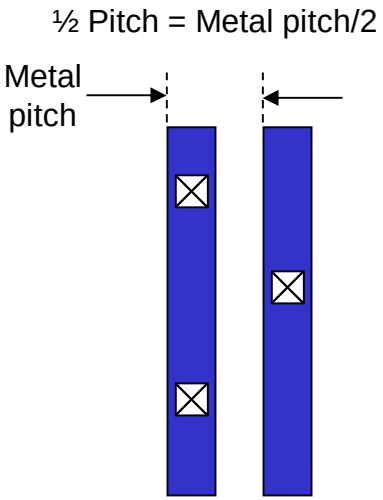
- International Technology Roadmap For Semiconductors (ITRS – 2012)
- Forecast from the semiconductor industry with a 15 year perspective:
  - Near-term: 2012 – 2019
  - Long-term: 2020 – 2026.
- The forecast is done in terms of 1st year of production:
  - Product shipment first exceeds 10K units/month (using production tooling)
- A near-term scaling ratio of @ 0.7 is assumed

In theory, there is no difference between theory and practice. But, in practice, there is.





Typical flash un-contacted poly



Typical DRAM/MPU/ASIC
Stagger-contact Metal Bit Line

2013 and beyond …

ITRS Road Map, 2012 edition:	2013	2017	2020	2023 (year of first production)
Gate length, physical (nm)	**20**	**14**	**11**	**8** (7 years for ½ L, 2 nodes)
Gate length, printed (nm)	28	17.7	12.5	8.8
Flash Poly ½ pitch (nm)	18	13	10	8 (flash drives the scaling)
DRAM ½ pitch (nm)	28	17..9	12.6	8.9
Wafer diameter (mm)	300	450	450	450 (wafers getting larger)
Wiring levels (maximum)	13	14	14	15 (tall metal stack)
DRAM:				
Bits/chip (Gbits)	4.29	8.59	34.36	*34.36* (increase and saturate)
Chip size (mm^2)	35	19	37	19 (small chips for low-cost bits)
Gbits/cm^2	12.24	46.25	92.5	185
Flash:				
Bits/chip (Gbits) SLC	69	137	275	412 (3D flash starts in 2016)
Chip size (mm^2)	149.8	135.88	162.42	158.13
Gbits/cm^2	45.9	101	169	261
Bits/chip (Gbits) MLC [3D 2 bits/cell]	-	550	1100	2199 (end of hard disks??)
Chip size (mm^2)	-	159.82	102.73	124.03
Gbits/cm^2	-	344	1070	1770
µP (high performance):				
Transistors/chip (Millions)	8848	17696	35391	70782 (doubling cores)
Chip size (mm^2)	260	206	206	206 (small chips for cheap MIPS)
Transistors/cm^2 (M/cm^2)	3403	8575	17150	34300
Total pads	3072	3072	3072	3072 (66.7% for power/ground)
Performance:				
On-Chip clock (GHz)	4	4.74	5.33	6 (saturate)
GFLOPs [Not ITRS data!] ~	100 (*)	237	533	1200 [not ITRS data!]
Power supply:				
Vdd (V):	0.85	0.75	0.68	0.62 (low-power for freq. increase)
Maximum allowable power (W)	149	130	130	? (With heat sink)