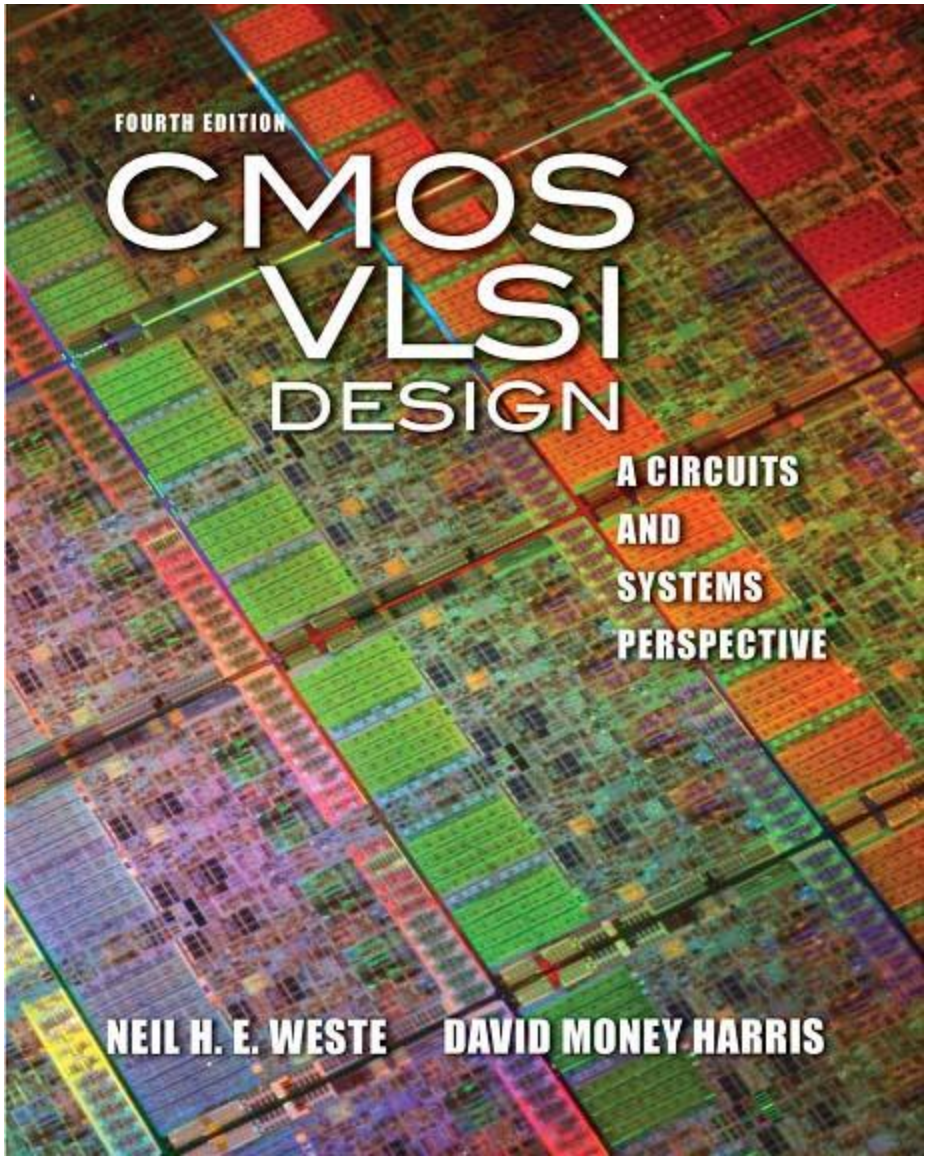


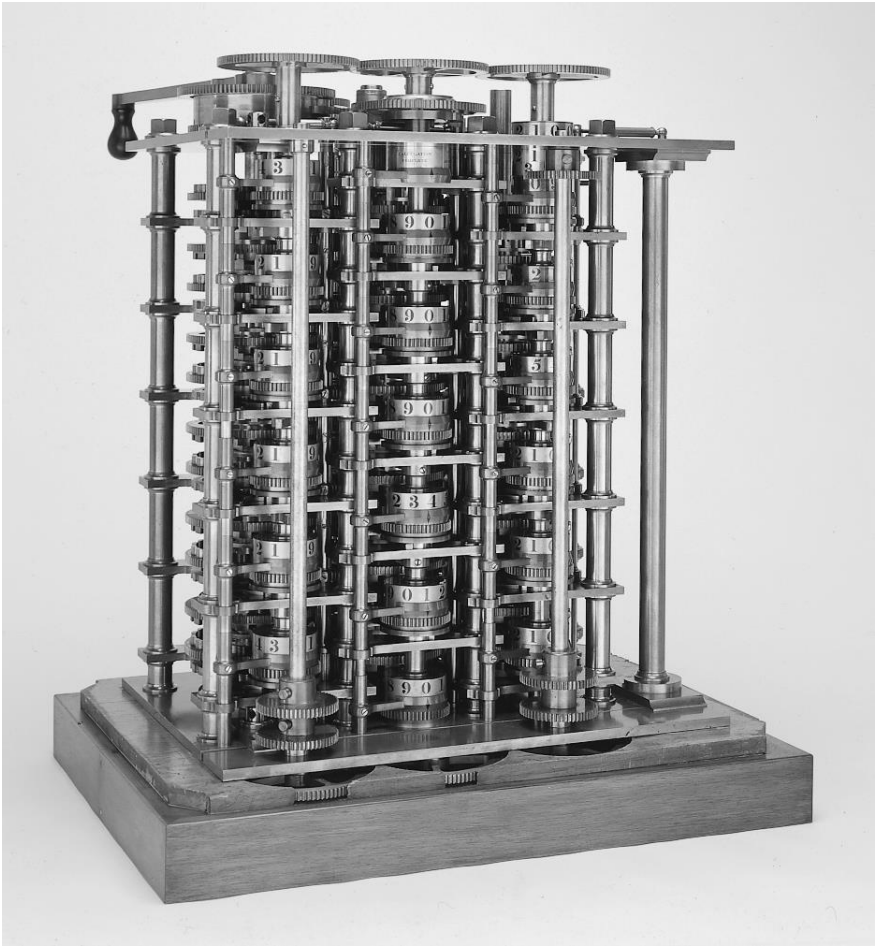
Lecture 2: Circuits & Layout

Outline

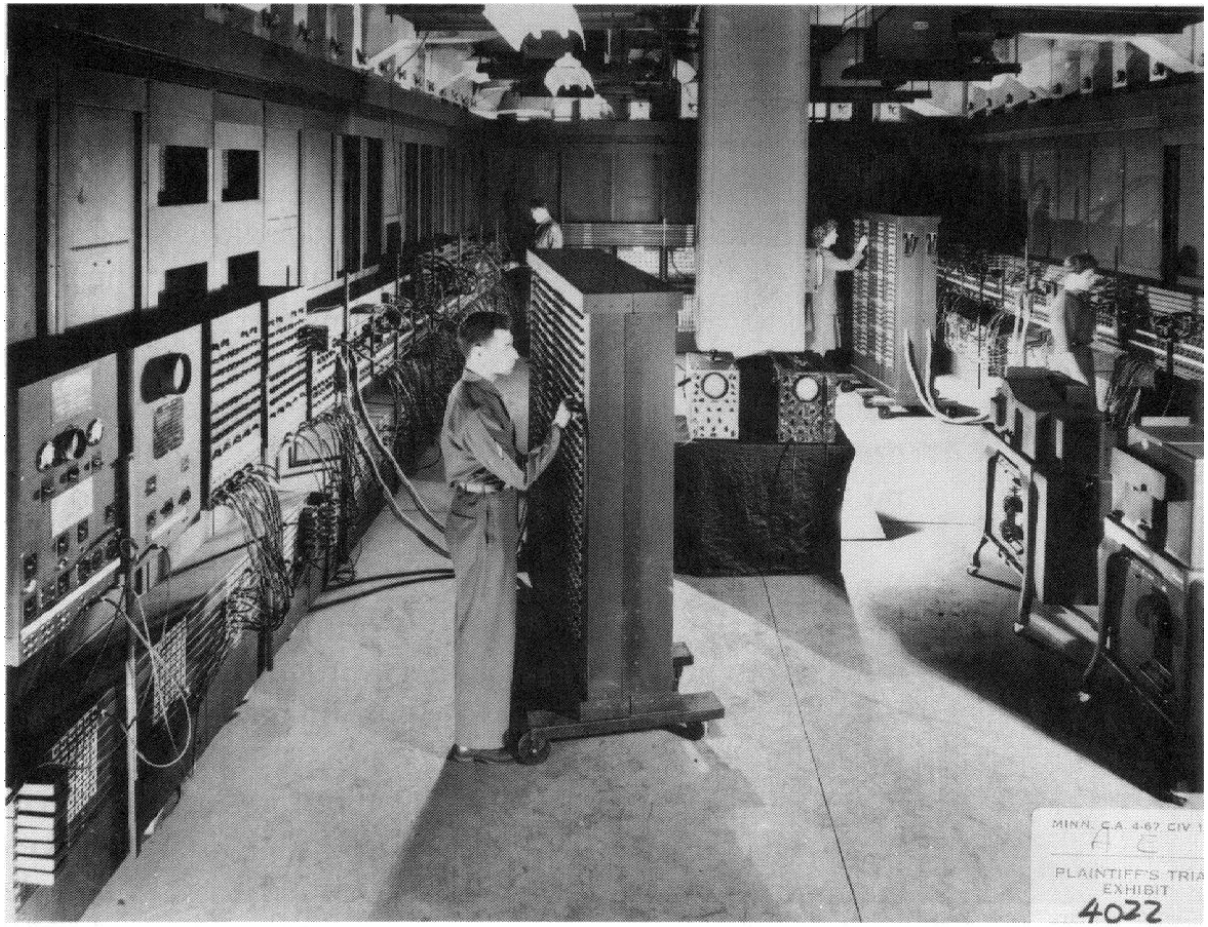
- A Brief History
- CMOS Gate Design
- Pass Transistors
- CMOS Latches & Flip-Flops
- Standard Cell Layouts
- Stick Diagrams

The First Computer

The Babbage Difference Engine (1832)
25.000 parts
cost: £17.470



ENIAC – The First Electronic Computer (1946)



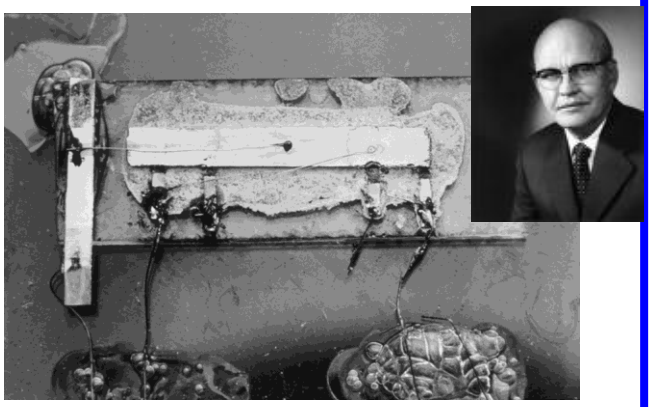
The Transistor Revolution

John Bardeen, Walter Brattain, and their supervisor William Shockley.

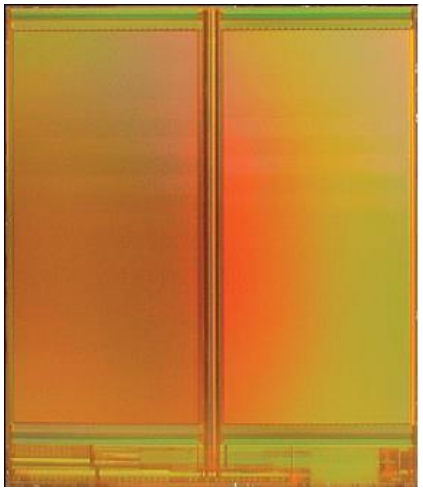
First Transistor
Bell Labs
1948

A Brief History

- 1958: First integrated circuit
  - Flip-flop using two transistors
  - Built by Jack Kilby at Texas Instruments
- 2010
  - Intel Core i7 μprocessor
    - 2.3 billion transistors
  - 64 Gb Flash memory
    - > 16 billion transistors

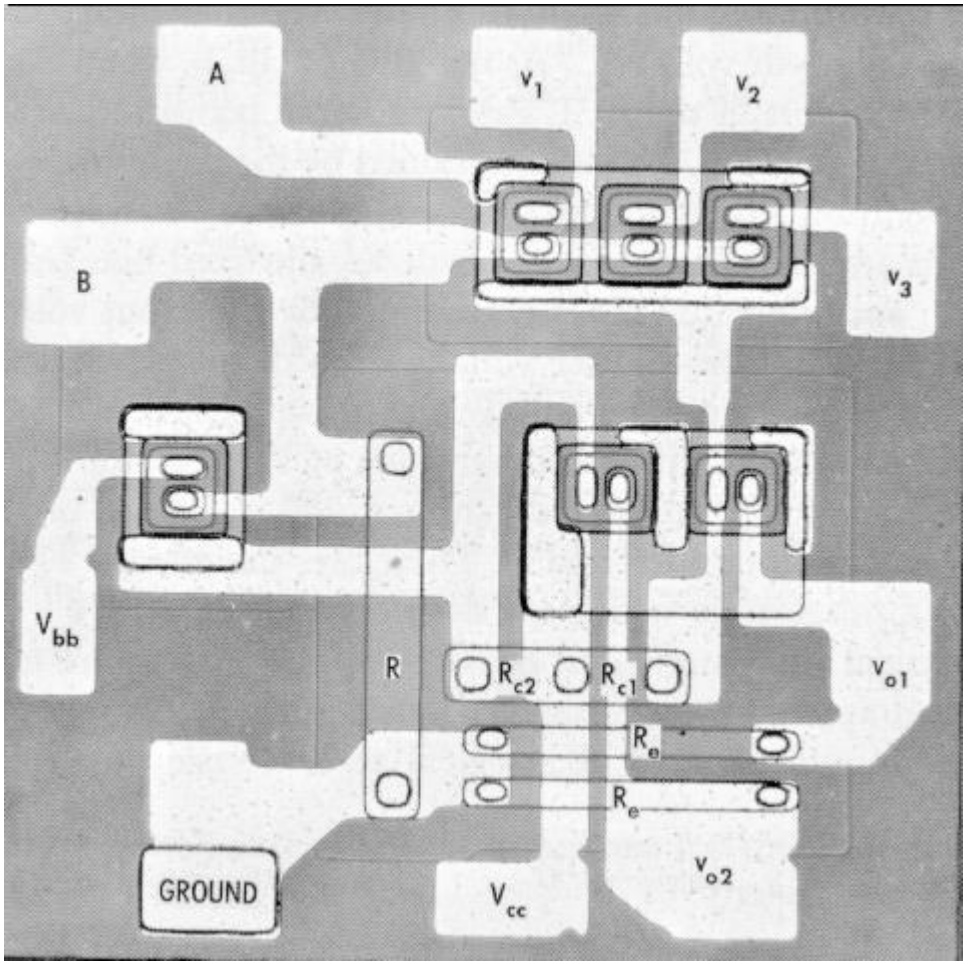


Courtesy Texas Instruments



[Trinh09]

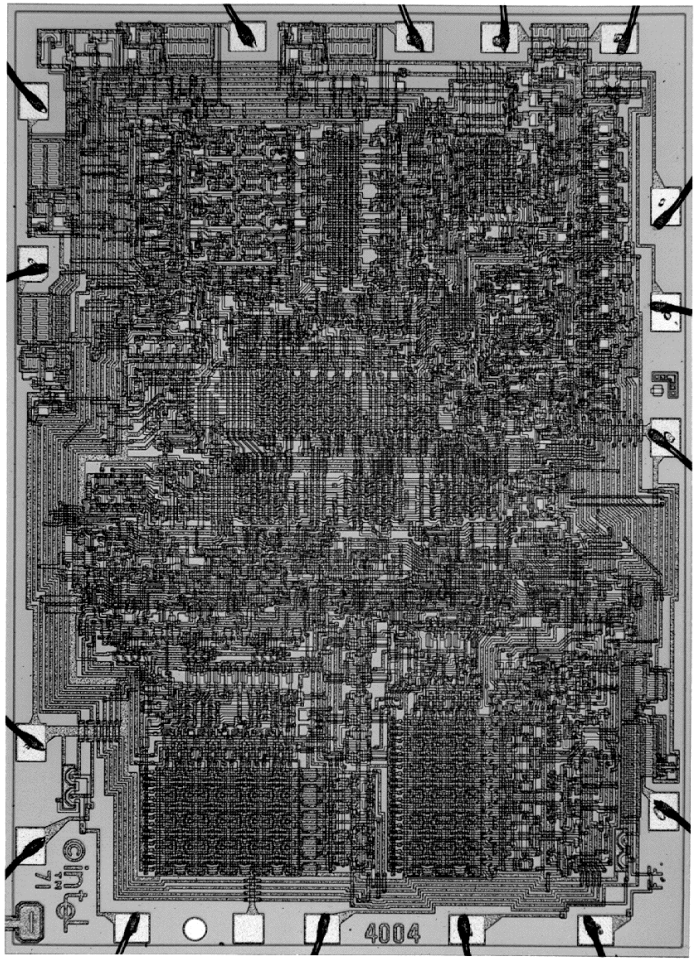
First Integrated Circuits



Bipolar Logic
1960's

ECL 3-input
NAND Gate
Motorola

Intel 4004 Microprocessor



1971
1000 transistors
1 MHz operation

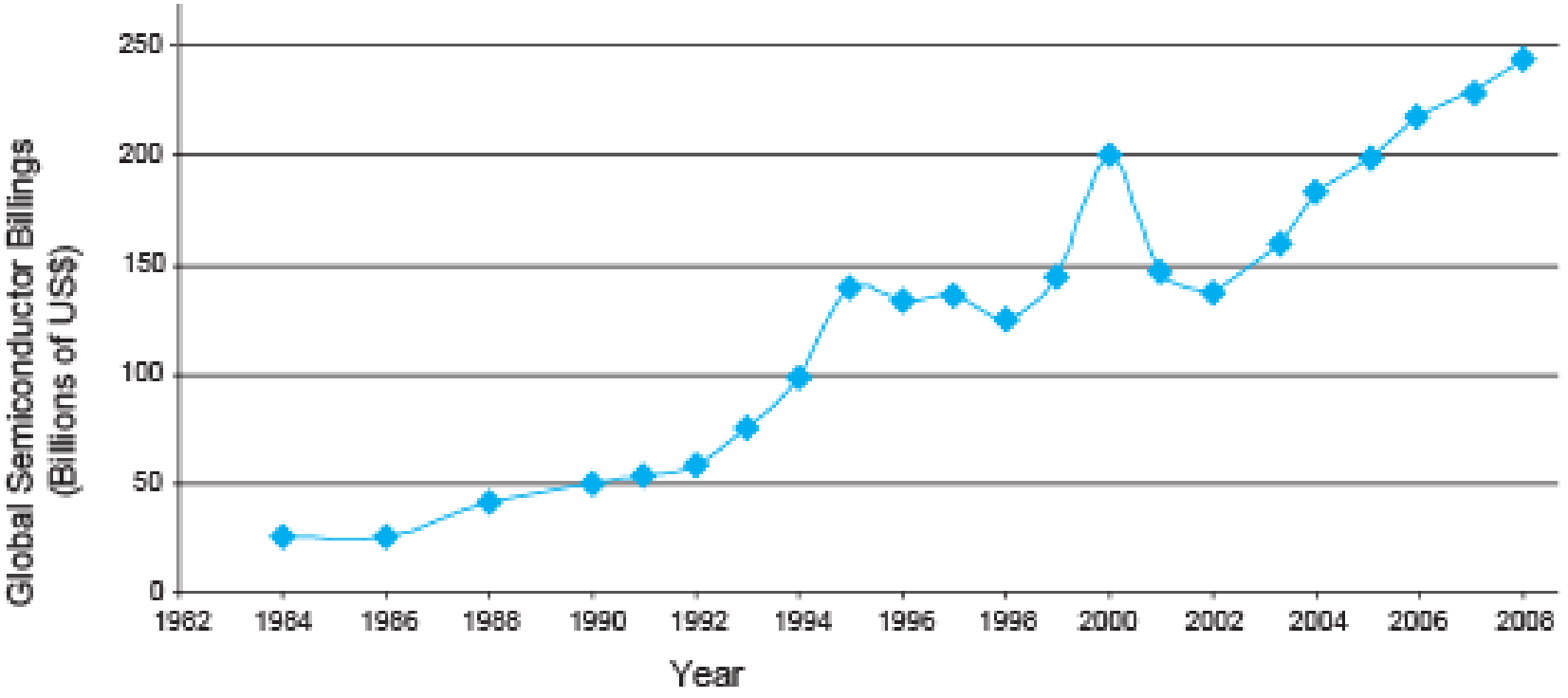
Growth Rate

- 53% compound annual growth rate over 50 years
  - No other technology has grown so fast so long
- Driven by miniaturization of transistors
  - Smaller is cheaper, faster, lower in power!
  - Revolutionary effects on society

[Moore65]
Electronics Magazine

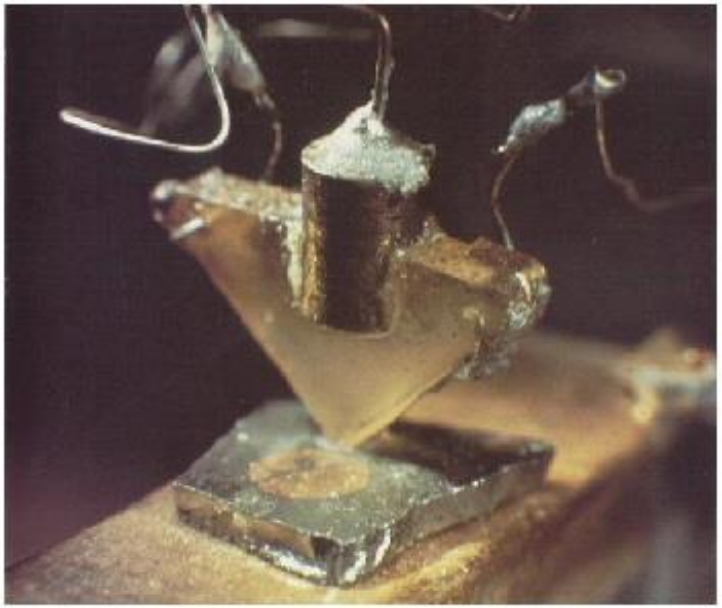
Annual Sales

- >10^{19} transistors manufactured in 2008
  - 1 billion for every human on the planet



Invention of the Transistor

- Vacuum tubes ruled in first half of 20th century
  Large, expensive, power-hungry, unreliable
- 1947: first point contact transistor
  - John Bardeen and Walter Brattain at Bell Labs
  - See *Crystal Fire*
    by Riordan, Hoddeson



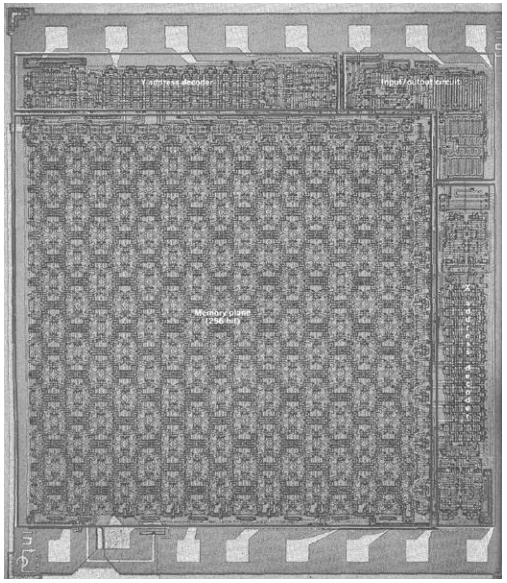
AT&T Archives.
Reprinted with
permission.

Transistor Types

- Bipolar transistors
  - npn or pnp silicon structure
  - Small current into very thin base layer controls large currents between emitter and collector
  - Base currents limit integration density
- Metal Oxide Semiconductor Field Effect Transistors
  - nMOS and pMOS MOSFETS
  - Voltage applied to insulated gate controls current between source and drain
  - Low power allows very high integration

MOS Integrated Circuits

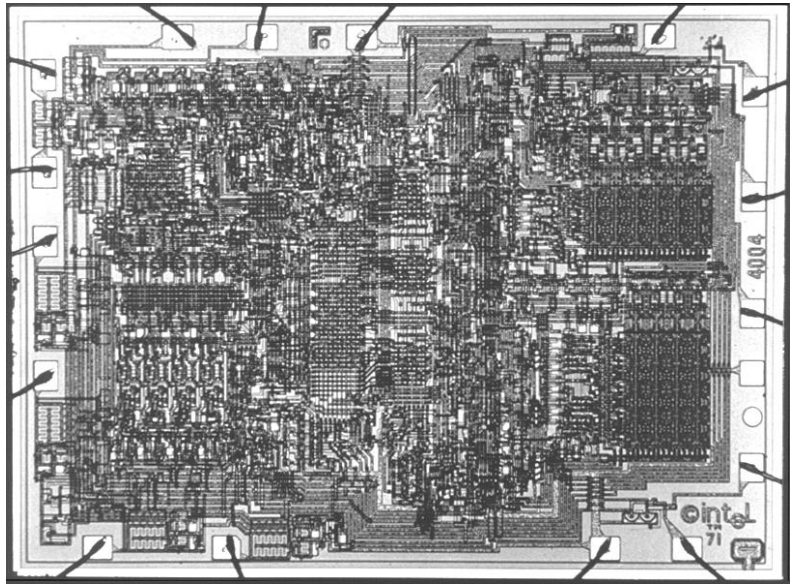
- 1970's processes usually had only nMOS transistors
  - Inexpensive, but consume power while idle



[Vadasz69]

© 1969 IEEE.

Intel 1101 256-bit SRAM



Intel Museum.

Reprinted with permission.

Intel 4004 4-bit μProc

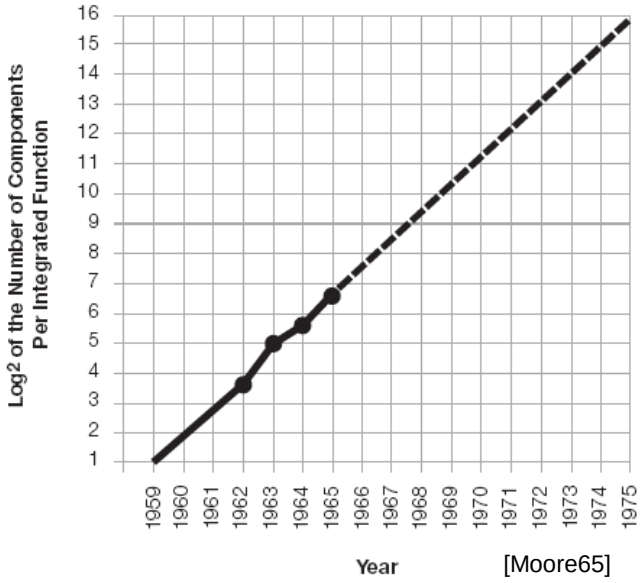
- 1980s-present: CMOS processes for low idle power

Moore's Law

- In 1965, Gordon Moore noted that the number of transistors on a chip doubled every 18 to 24 months.
- He made a prediction that semiconductor technology will double its effectiveness every 18 months

Moore's Law: Then

- 1965: Gordon Moore plotted transistor on each chip
  - Fit straight line on semilog scale
  - Transistor counts have doubled every 26 months



[Moore65]

Electronics Magazine

Integration Levels

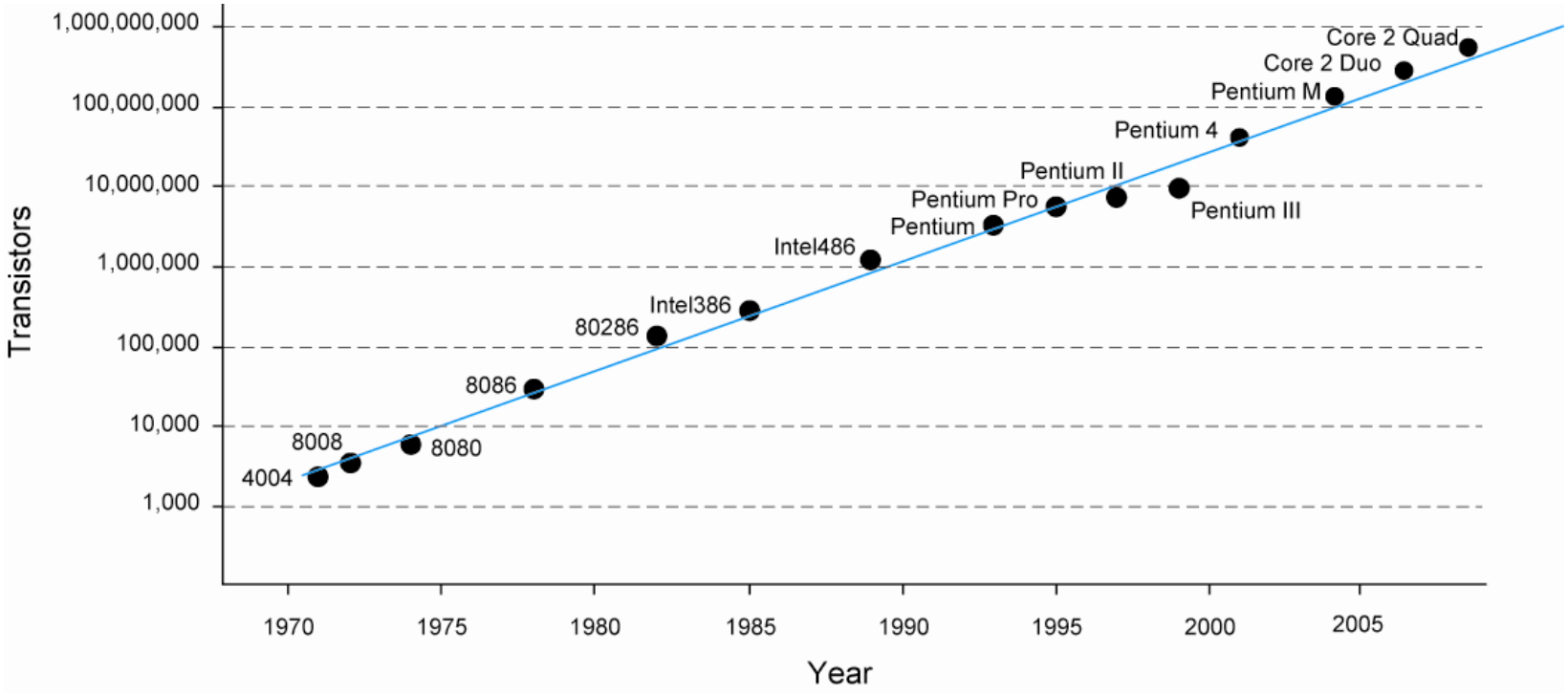
SSI: 10 gates

MSI: 1000 gates

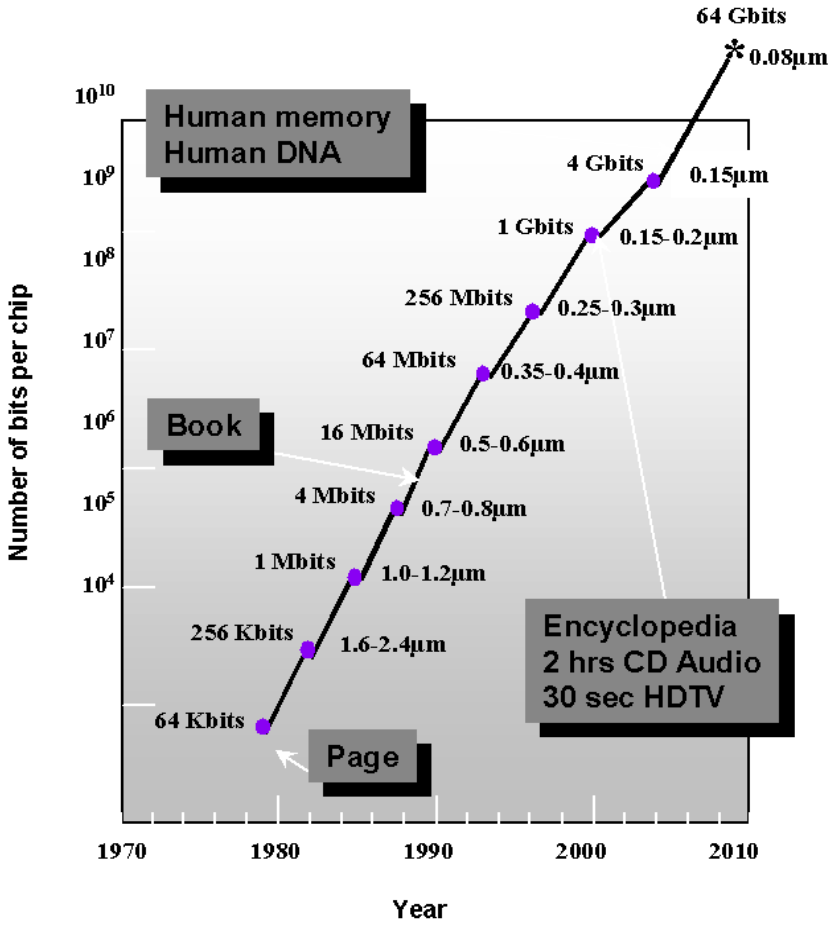
LSI: 10,000 gates

VLSI: > 10k gates

And Now…


1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
Transistors
1970
1975
1980
1985
1990
1995
2000
2005
Year
4004
8008
8080
8086
80286
Intel386
Intel486
Pentium
Pentium Pro
Pentium II
Pentium III
Pentium 4
Pentium M
Core 2 Duo
Core 2 Quad

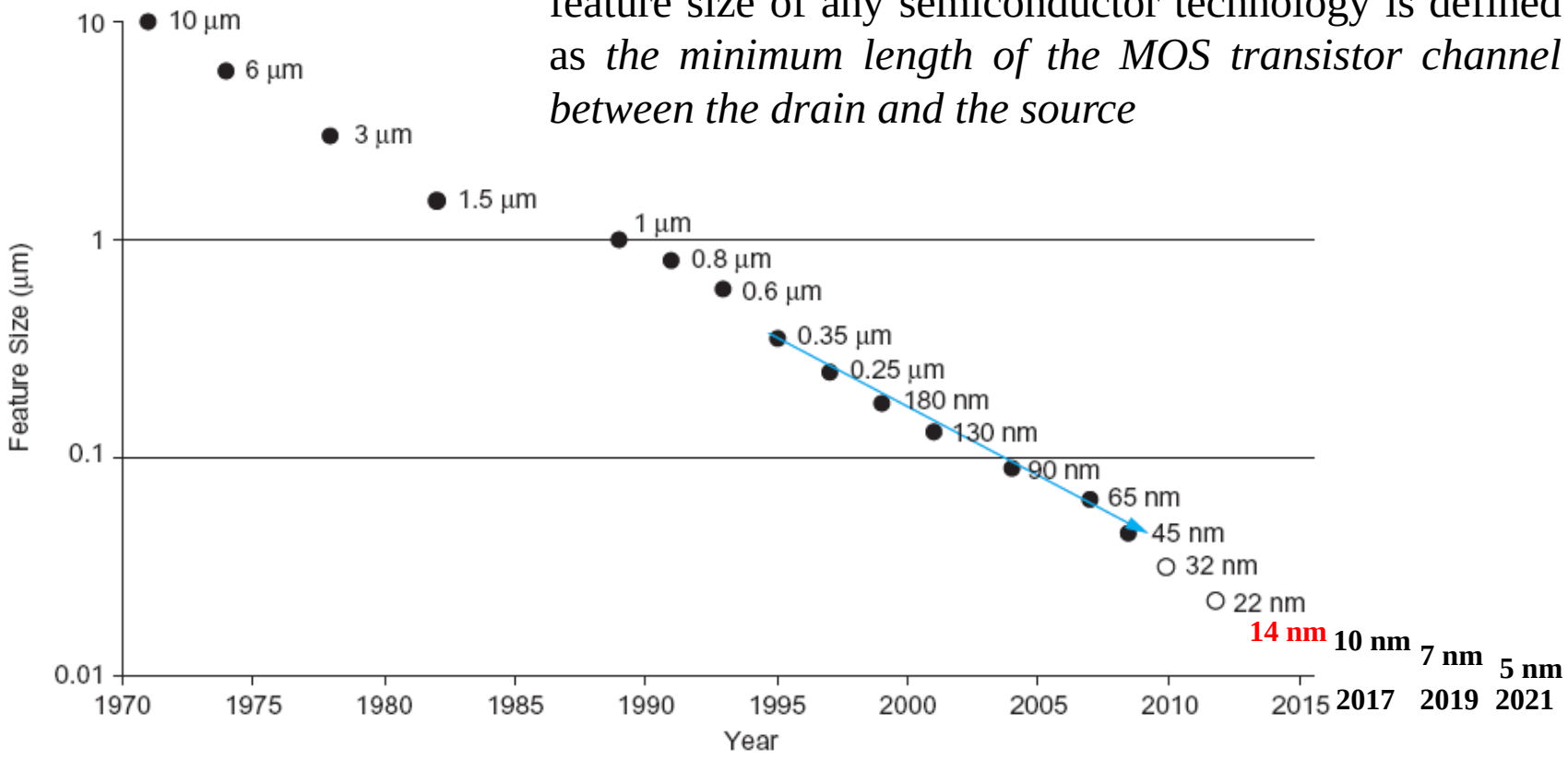
Evolution in Complexity



Feature Size

- Minimum feature size shrinking 30% every 2-3 years

feature size of any semiconductor technology is defined as *the minimum length of the MOS transistor channel between the drain and the source*

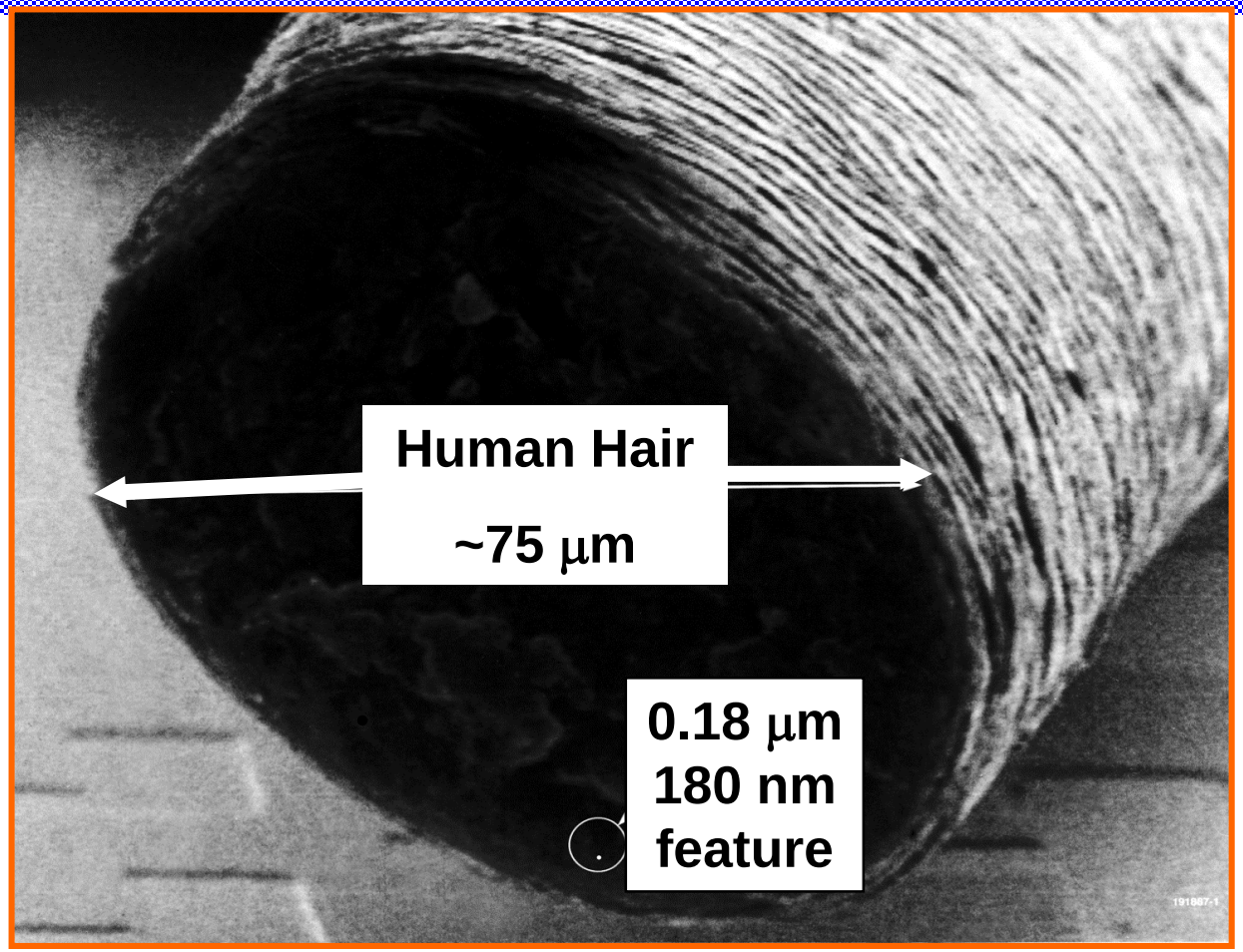


Feature Size

TABLE 7.6 Predictions from the 2007 ITRS

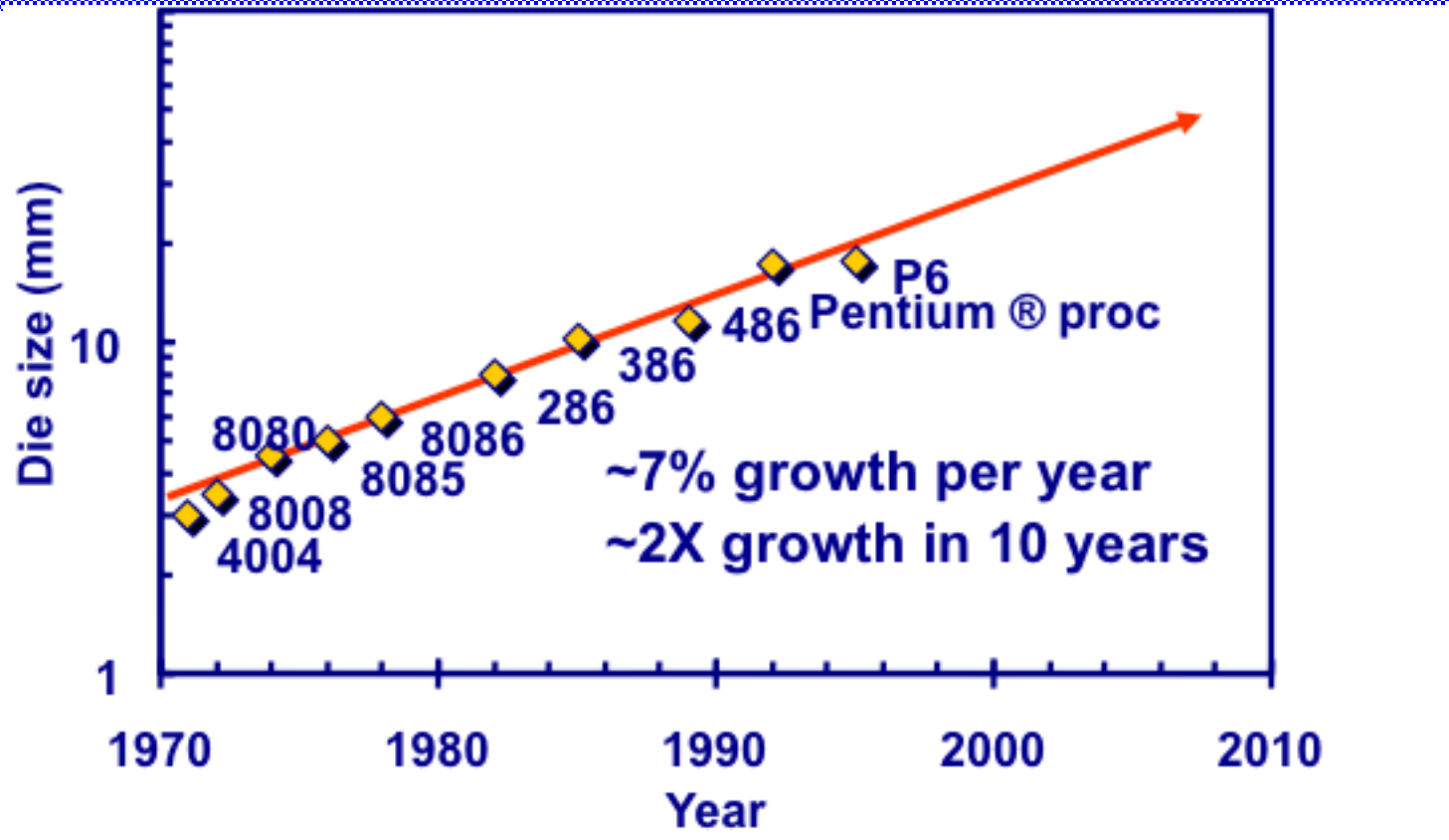
Year	2009	2012	2015	2018	2021
Feature size (nm)	34	24	17	12	8.4
L_{gate} (nm)	20	14	10	7	5
V_{DD} (V)	1.0	0.9	0.8	0.7	0.65
Billions of transistors/die	1.5	3.1	6.2	12.4	24.7
Wiring levels	12	12	13	14	15
Maximum power (W)	198	198	198	198	198
DRAM capacity (Gb)	2	4	8	16	32
Flash capacity (Gb)	16	32	64	128	256

Feature Size



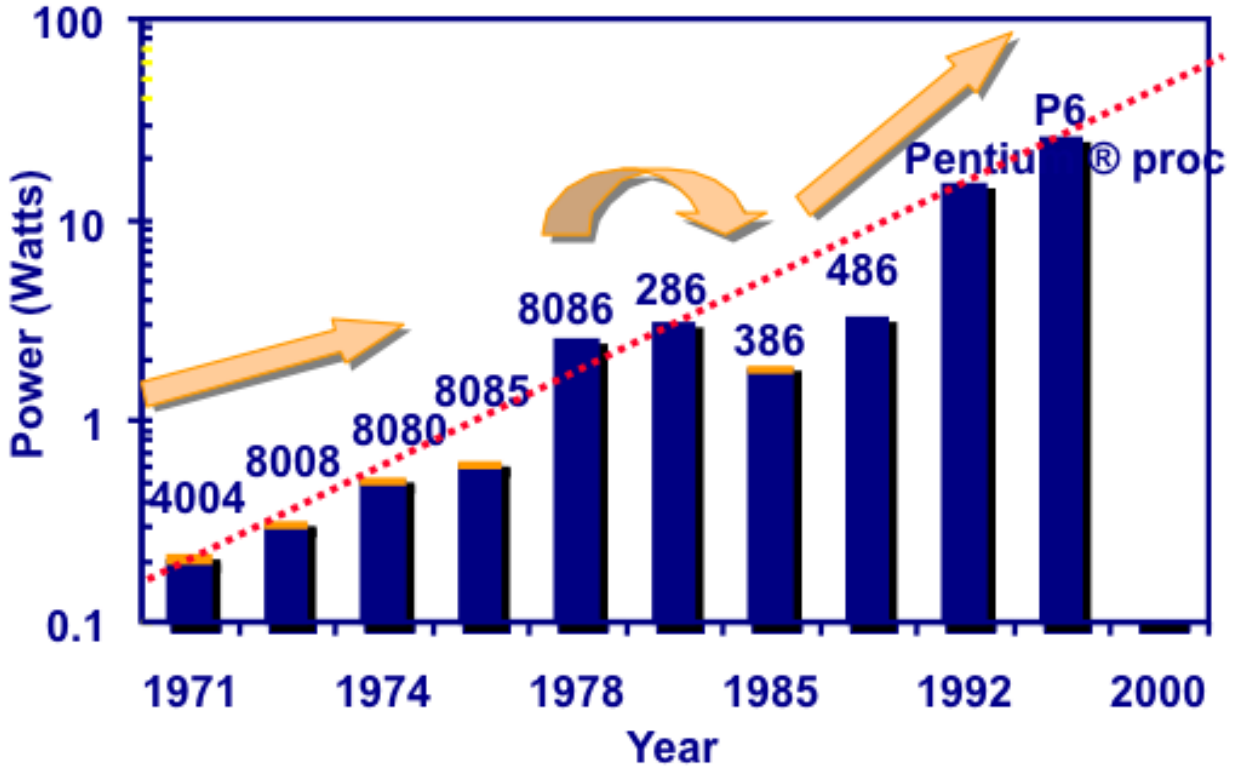
~40,000 (65-nm node) transistors could fit on cross-section

Die Size Growth



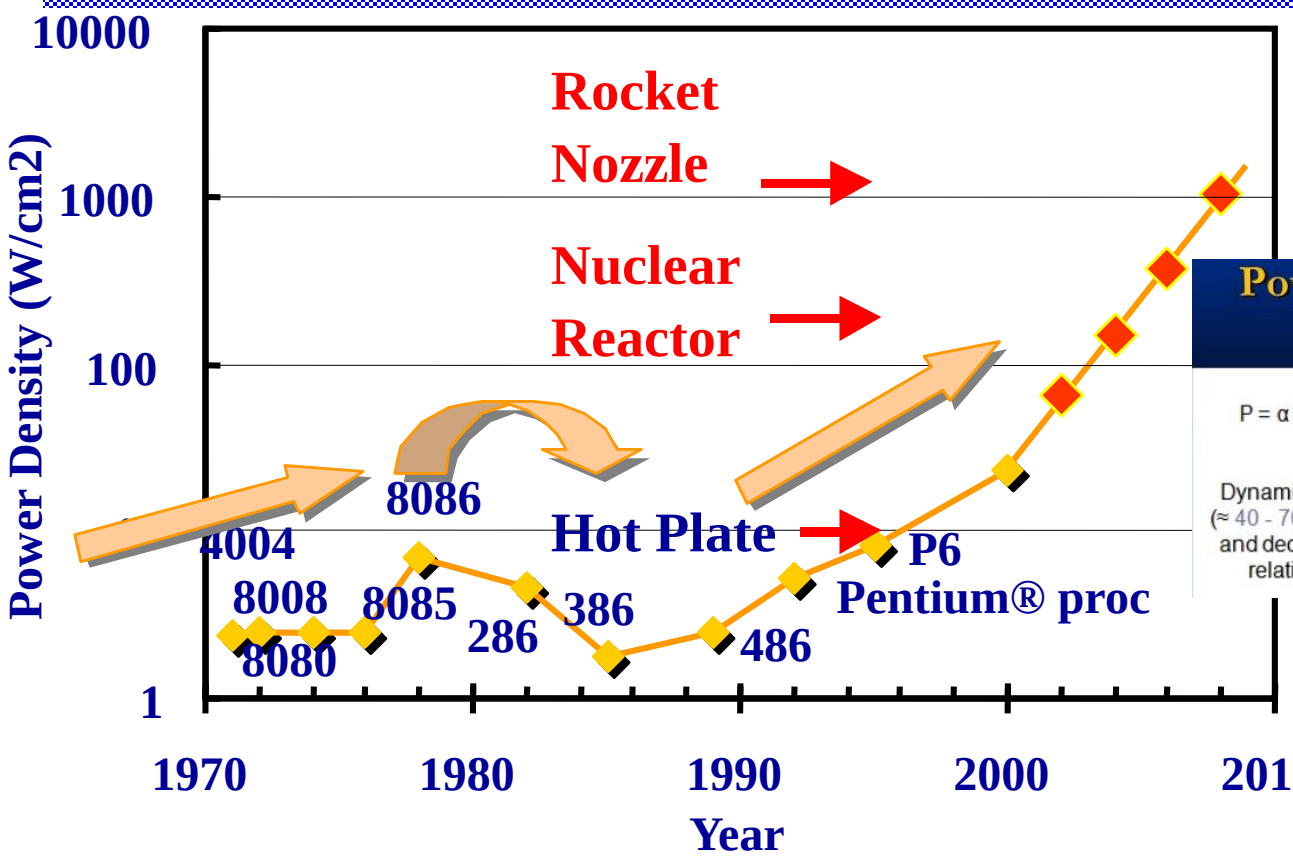
Die size grows by 14% to satisfy Moore's Law

Power Dissipation



Lead Microprocessors power continues to increase

Power density



Power Equations in CMOS

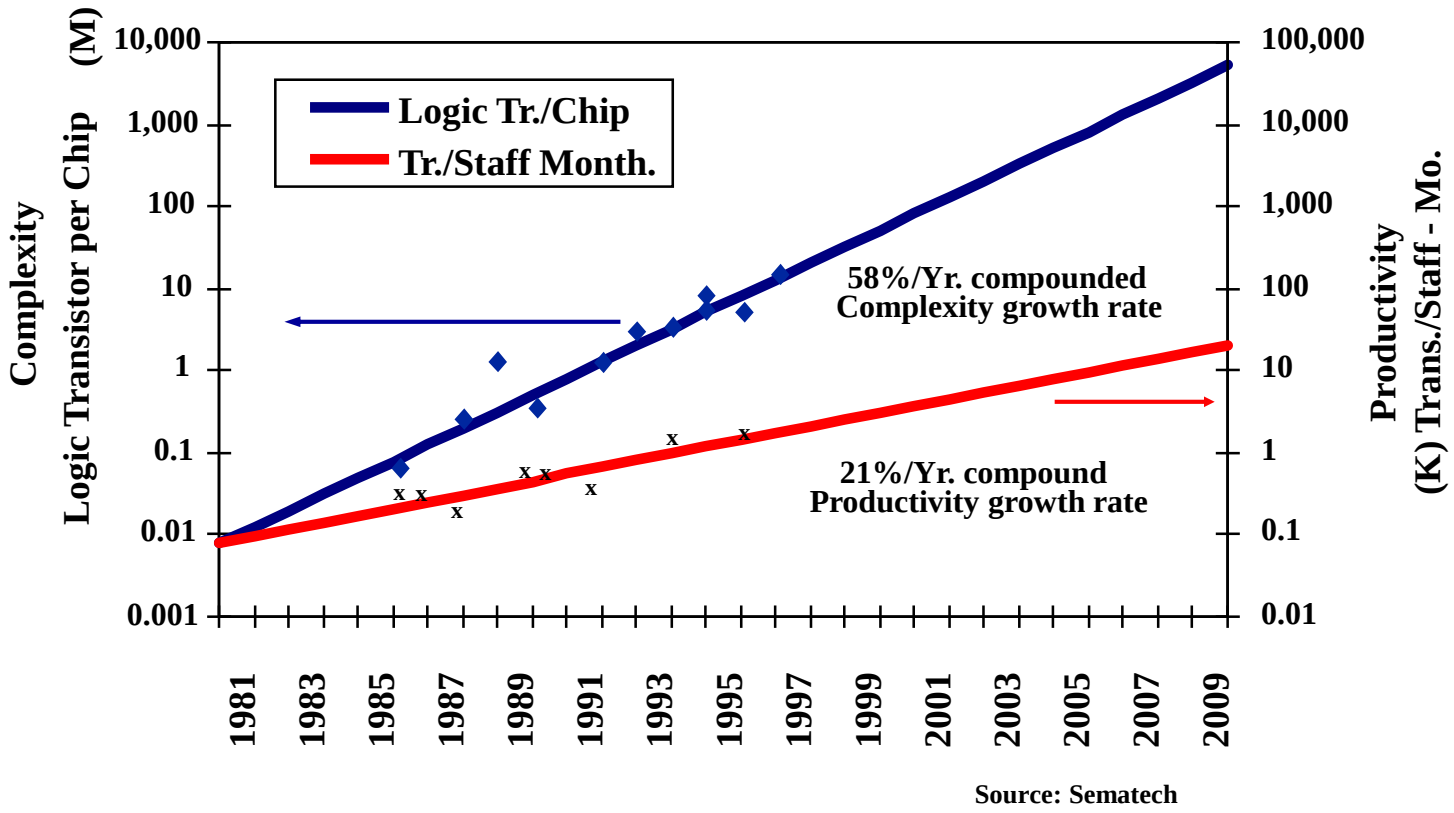
$$P = \alpha f C_L V_{DD}^2 + V_{DD} I_{peak} (P_{0\to1} + P_{1\to0}) + V_{DD} I_{leak}$$

Dynamic power	Short-circuit power	Leakage power
(≈ 40 - 70% today and decreasing relatively)	(≈ 10 % today and decreasing absolutely)	(≈ 20 – 50 % today and increasing)

Power density too high to keep junctions at low temp

Courtesy, Intel

Productivity Trends



Source: Sematech

Complexity outpaces design productivity

Courtesy, ITRS Roadmap

Why Scaling?

- Technology shrinks by 0.7/generation
- With every generation can integrate 2x more functions per chip; chip cost does not increase significantly
- Cost of a function decreases by 2x
- But …
  - How to design chips with more and more functions?
  - Design engineering population does not double every two years…
- Hence, a need for more efficient design methods
  - Exploit different levels of abstraction

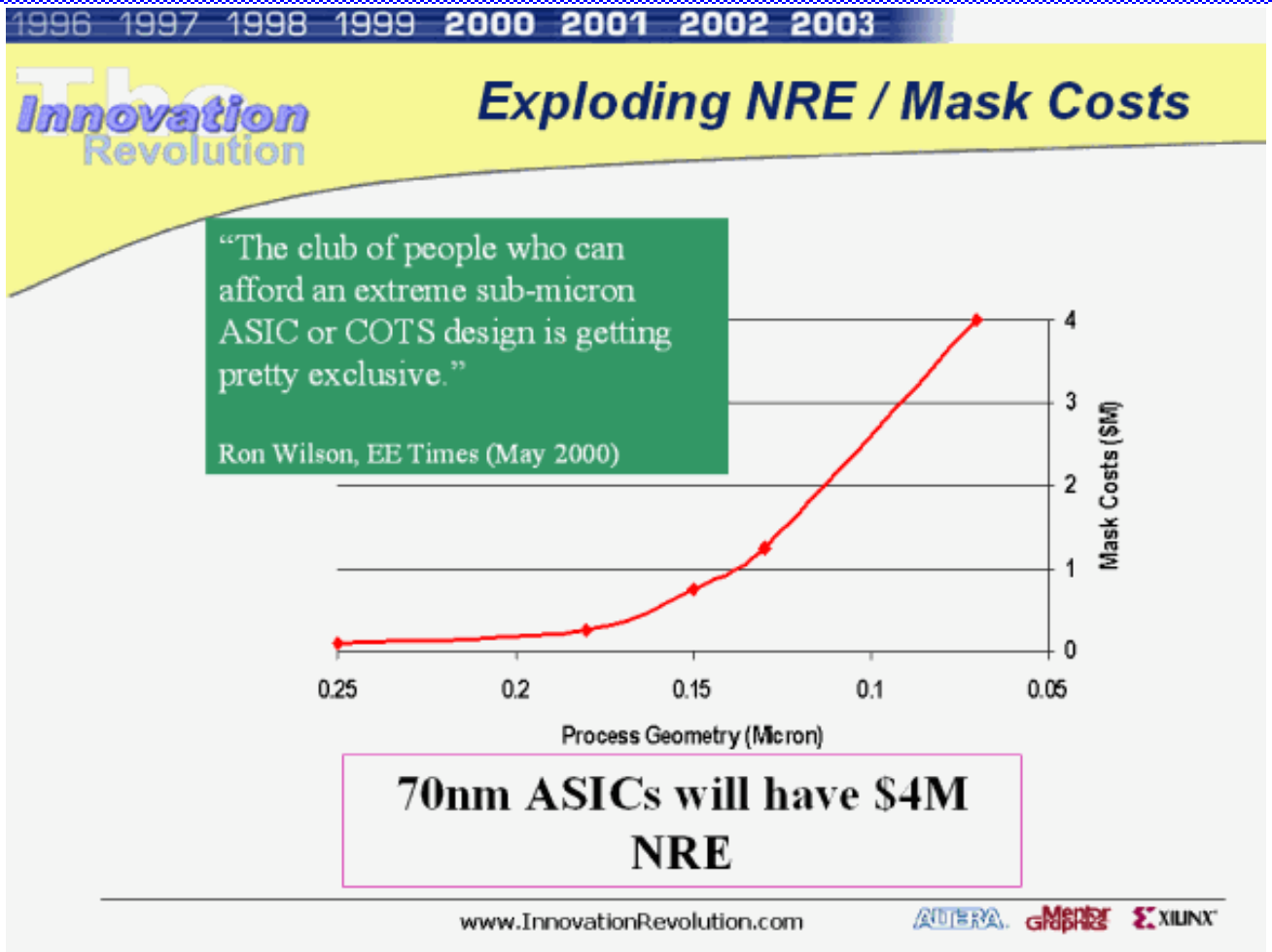
Design Metrics

- How to evaluate performance of a digital circuit (gate, block, …)?
  - Cost
  - Reliability
  - Scalability
  - Speed (delay, operating frequency)
  - Power dissipation
  - Energy to perform a function

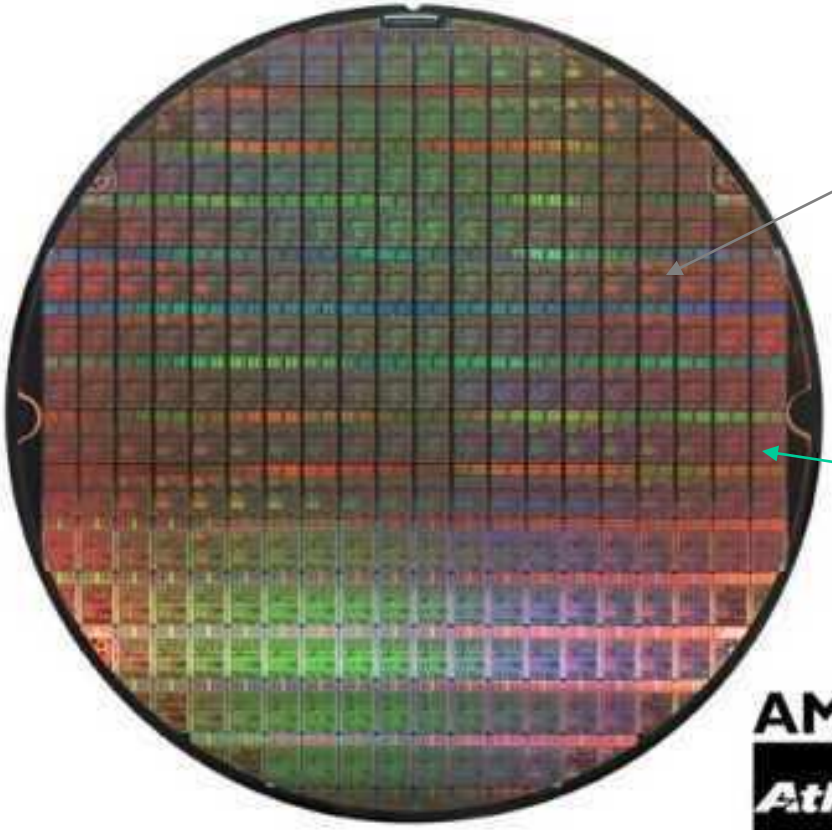
Cost of Integrated Circuits

- NRE (non-recurrent engineering) costs
  - design time and effort, mask generation
  - one-time cost factor
- Recurrent costs
  - silicon processing, packaging, test
  - proportional to volume
  - proportional to chip area

NRE Cost is Increasing



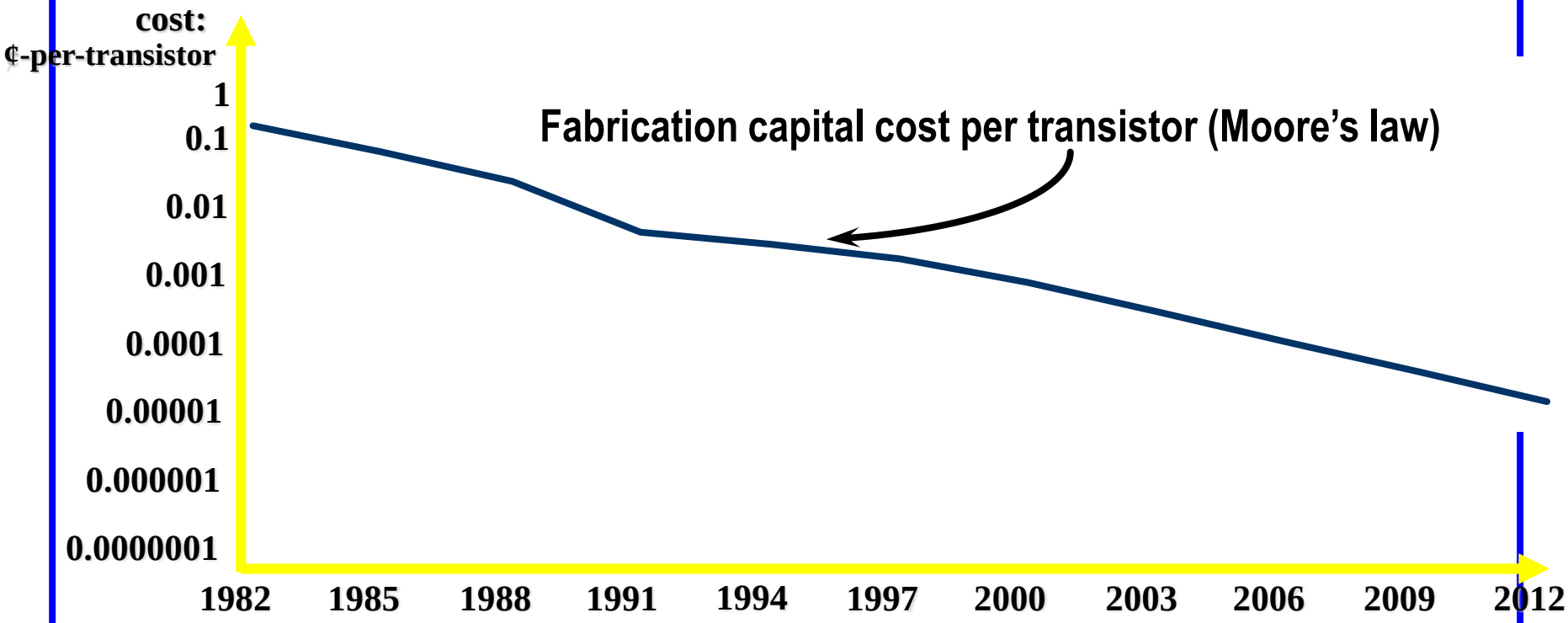
Die Cost



AMD Athlon

Going up to 12" (30cm)

Cost per Transistor

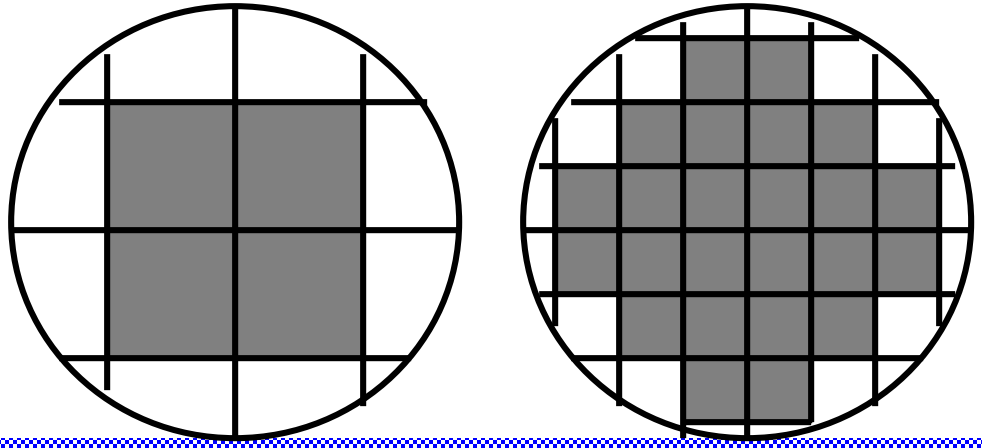


Yield

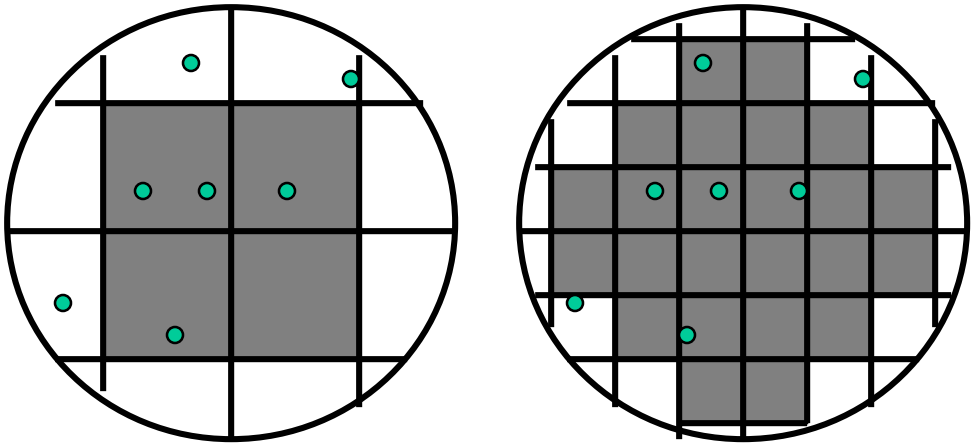
$$Y = \frac{\text{No. of good chips per wafer}}{\text{Total number of chips per wafer}} \times 100\%$$

$$\text{Die cost} = \frac{\text{Wafer cost}}{\text{Dies per wafer} \times \text{Die yield}}$$

$$\text{Dies per wafer} = \frac{\pi \times (\text{wafer diameter}/2)^2}{\text{die area}} - \frac{\pi \times \text{wafer diameter}}{\sqrt{2 \times \text{die area}}}$$



Defects



$$\text{die yield} = \left(1 + \frac{\text{defects per unit area} \times \text{die area}}{\alpha}\right)^{-\alpha}$$

α is approximately 3

$$\text{die cost} = f(\text{die area})^4$$

Some Examples (1994)

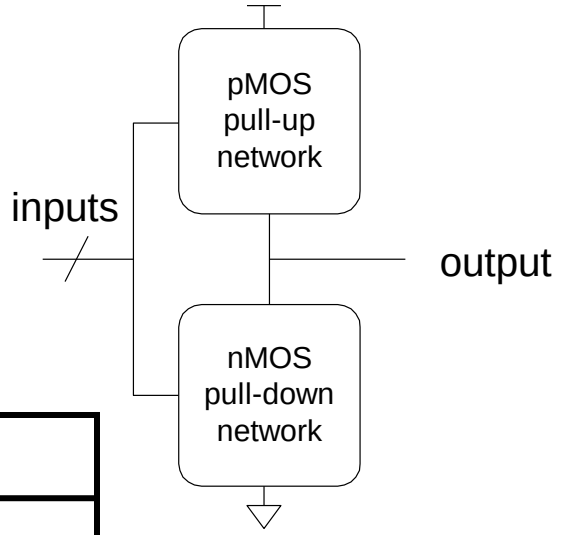
Chip	Metal layers	Line width	Wafer cost	Def./ cm^2	Area mm^2	Dies/ wafer	Yield	Die cost
386DX	2	0.90	$900	1.0	43	360	71%	$4
486 DX2	3	0.80	$1200	1.0	81	181	54%	$12
Power PC 601	4	0.80	$1700	1.3	121	115	28%	$53
HP PA 7100	3	0.80	$1300	1.0	196	66	27%	$73
DEC Alpha	3	0.70	$1500	1.2	234	53	19%	$149
Super Sparc	3	0.70	$1700	1.6	256	48	13%	$272
Pentium	3	0.80	$1500	1.5	296	40	9%	$417

CMOS Gate Design

- Activity:
  - Sketch a 4-input CMOS NOR gate

Complementary CMOS

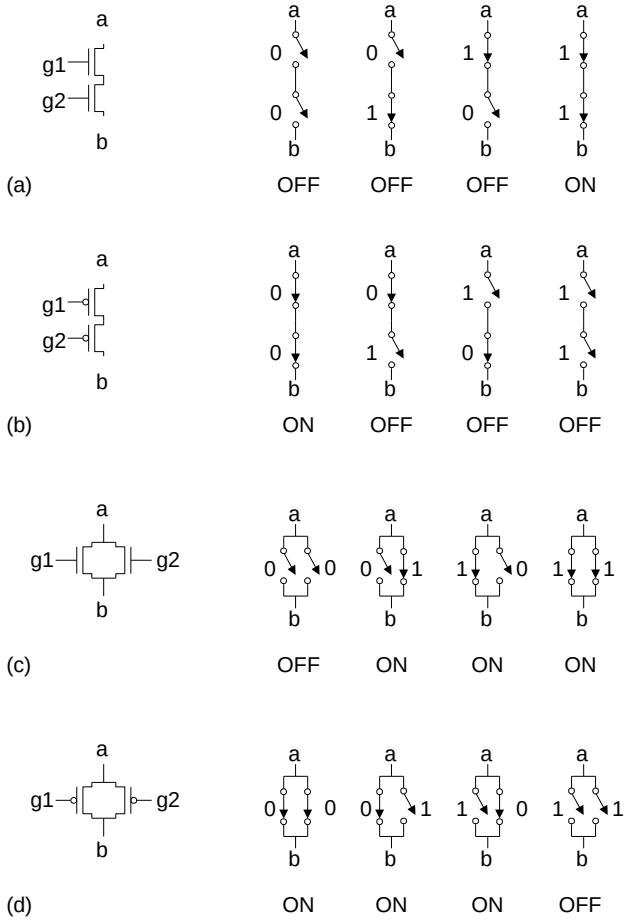
- Complementary CMOS logic gates
  - nMOS *pull-down network*
  - pMOS *pull-up network*
  - a.k.a. static CMOS



	Pull-up OFF	Pull-up ON
Pull-down OFF	Z (float)	1
Pull-down ON	0	X (crowbar)

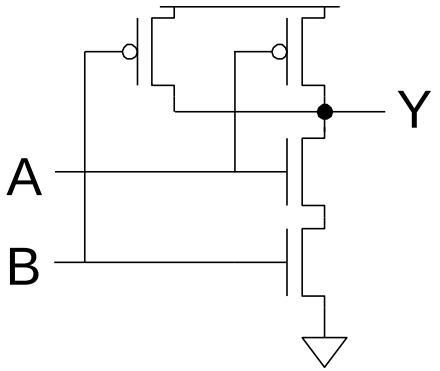
Series and Parallel

- nMOS: 1 = ON
- pMOS: 0 = ON
- *Series*: both must be ON
- *Parallel*: either can be ON



Conduction Complement

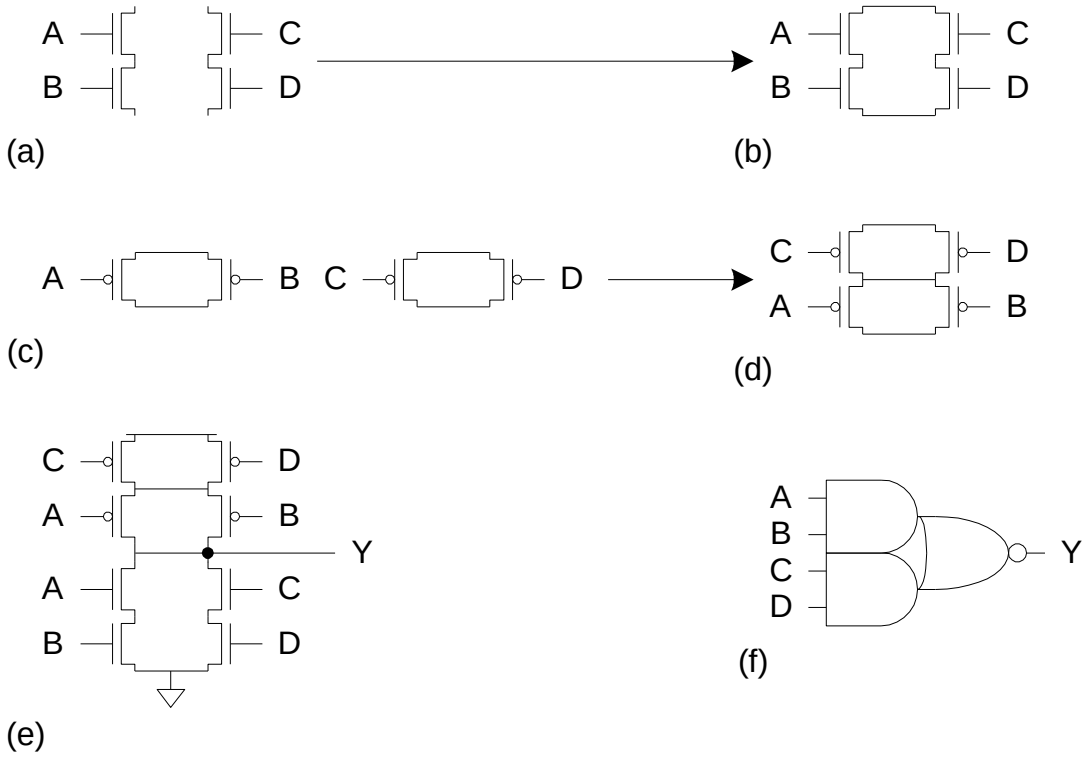
- Complementary CMOS gates always produce 0 or 1
- Ex: NAND gate
  - Series nMOS: Y=0 when both inputs are 1
  - Thus Y=1 when either input is 0
  - Requires parallel pMOS



- Rule of *Conduction Complements*
  - Pull-up network is complement of pull-down
  - Parallel -> series, series -> parallel

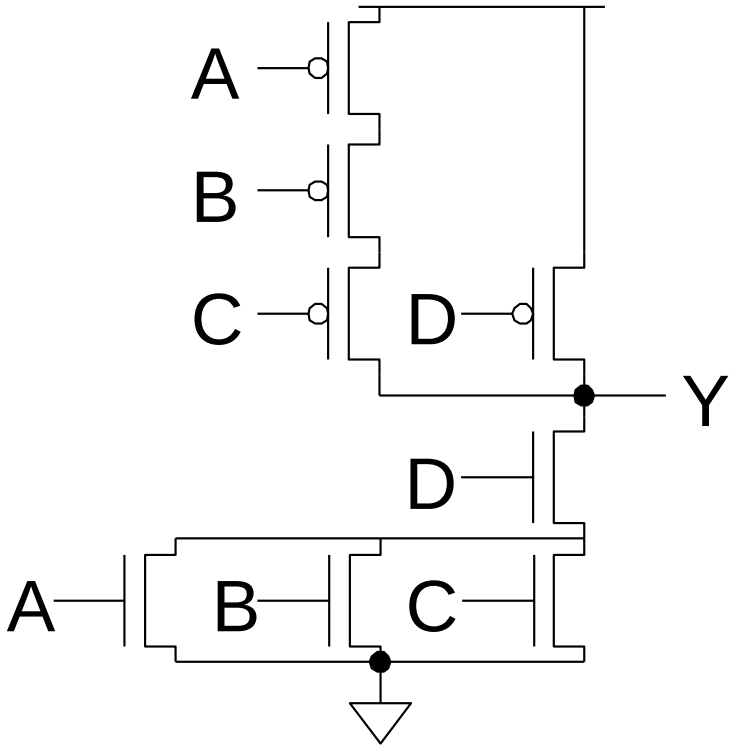
Compound Gates

- *Compound gates* can do any inverting function
- Ex: $Y = \overline{A \cdot B + C \cdot D}$ (AND-AND-OR-INVERT, AOI22)



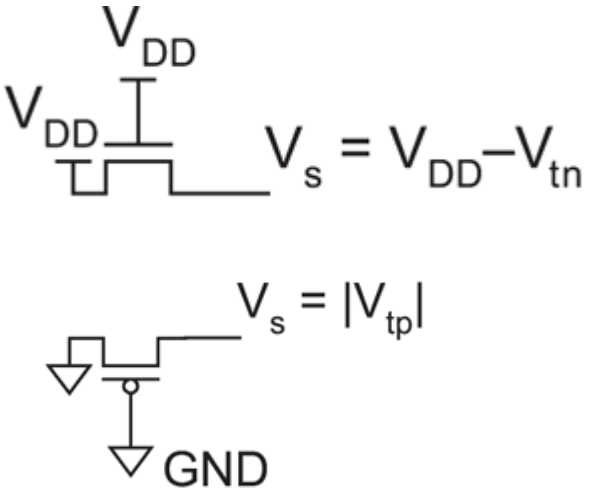
Example: O3AI

- $Y = \overline{(A+B+C)\cdot D}$



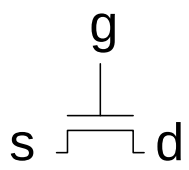
Signal Strength

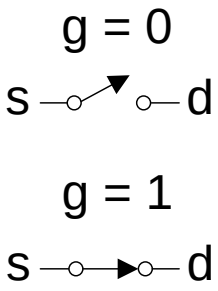
- *Strength* of signal
  - How close it approximates ideal voltage source
- V_{DD} and GND rails are strongest 1 and 0
- nMOS pass strong 0
  - But degraded or weak 1
- pMOS pass strong 1
  - But degraded or weak 0
- Thus nMOS are best for pull-down network

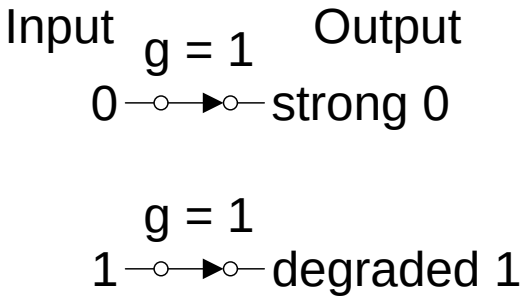


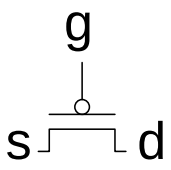
Pass Transistors

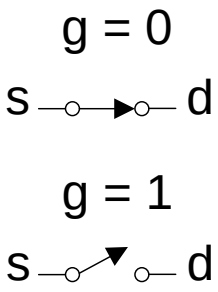
- Transistors can be used as switches

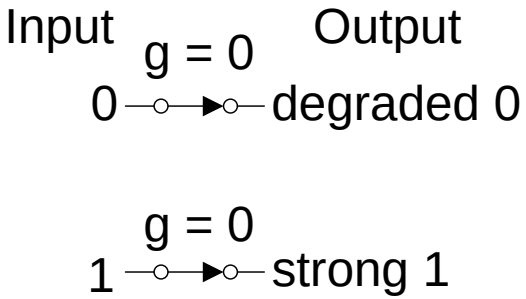






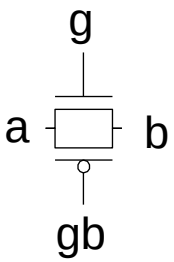




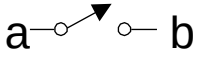


Transmission Gates

- Pass transistors produce degraded outputs
- *Transmission gates* pass both 0 and 1 well

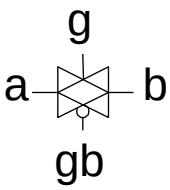


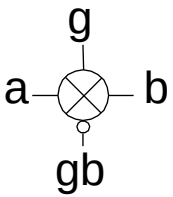
g = 0, gb = 1

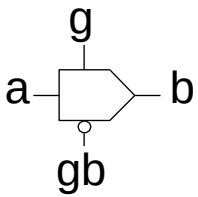


g = 1, gb = 0

Input	Output
g = 1, gb = 0 0	strong 0
g = 1, gb = 0 1	strong 1



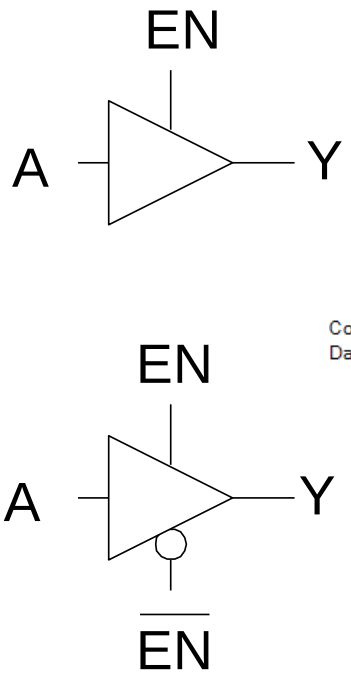


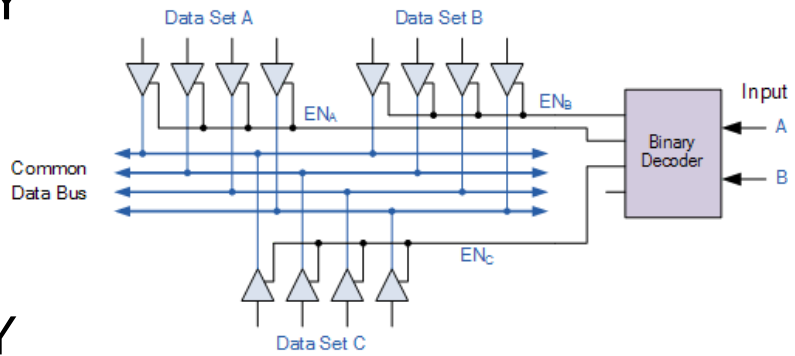


Tristates

- *Tristate buffer* produces Z when not enabled

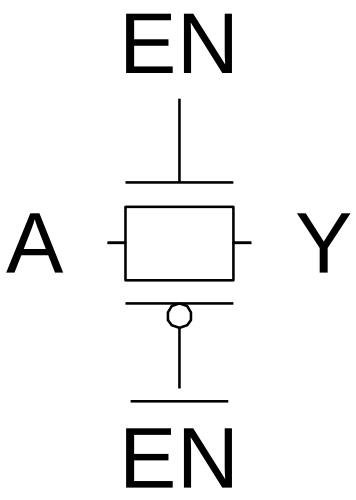
EN	A	Y
0	0	
0	1	
1	0	
1	1	





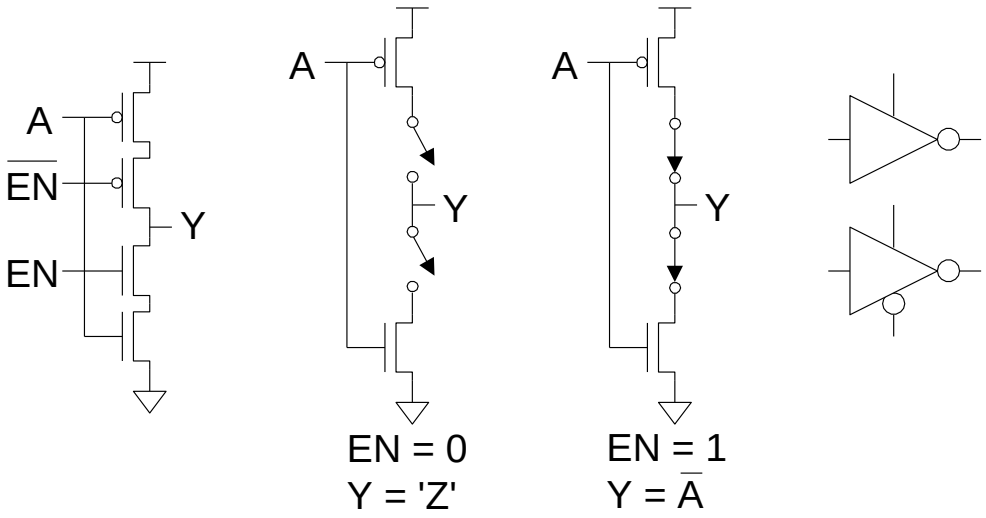
Nonrestoring Tristate

- Transmission gate acts as tristate buffer
  - Only two transistors
  - But *nonrestoring*
    - Noise on A is passed on to Y



Tristate Inverter

- Tristate inverter produces restored output
  - Violates conduction complement rule
  - Because we want a Z output



Tristates were once commonly used to allow multiple units to drive a common bus, as long as exactly one unit is enabled at a time.

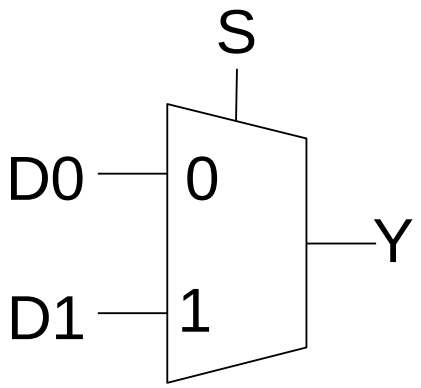
- If multiple units drive the bus, contention occurs and power is wasted.
- If no units drive the bus, it can float to an invalid logic level that causes the receivers to waste power.
- Moreover, it can be difficult to switch enable signals at exactly the same time when they are distributed across a large chip. Delay between different enables switching can cause contention.

Given these problems, multiplexers are now preferred over tristate busses

Multiplexers

- 2:1 multiplexer chooses between two inputs

S	D1	D0	Y
0	X	0	
0	X	1	
1	0	X	
1	1	X	



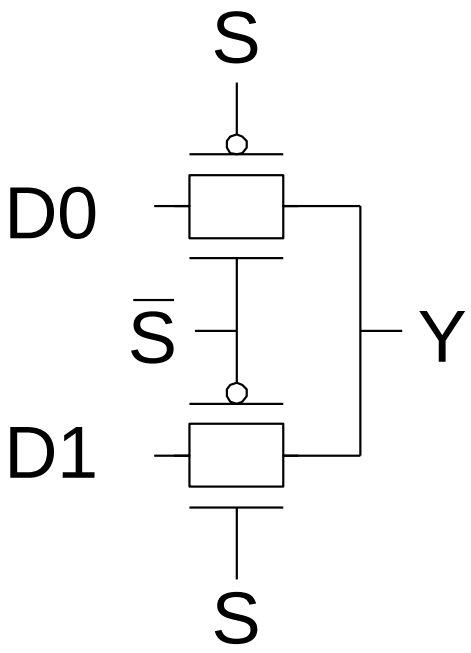
chooses input D0 when the select is 0 and input D1 when the select is 1.

Gate-Level Mux Design

- $Y = SD_1 + \overline{S}D_0$ (too many transistors)
- How many transistors are needed?

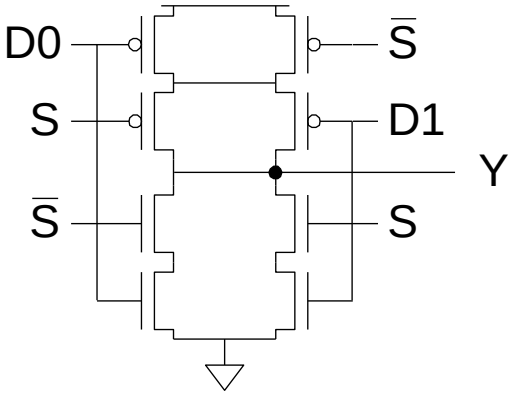
Transmission Gate Mux

- Nonrestoring mux uses two transmission gates
  - Only 4 transistors

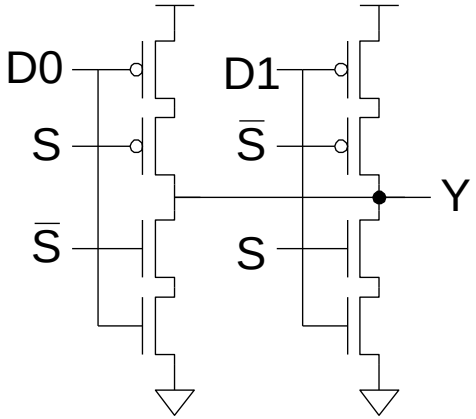


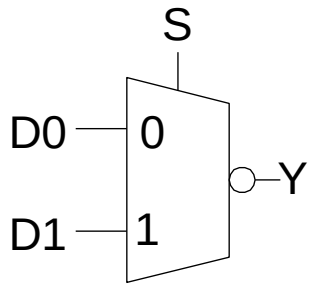
Inverting Mux

- Inverting multiplexer
  - Use compound AOI22
  - Or pair of tristate inverters
  - Essentially the same thing
- Noninverting multiplexer adds an inverter



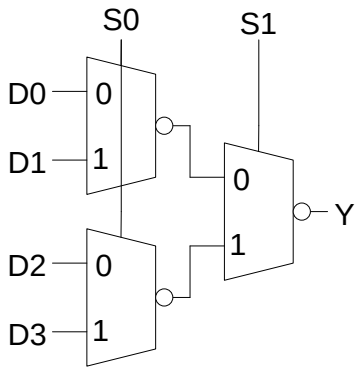
$$\overline{y = SD_1 + \bar{S}D_0}$$

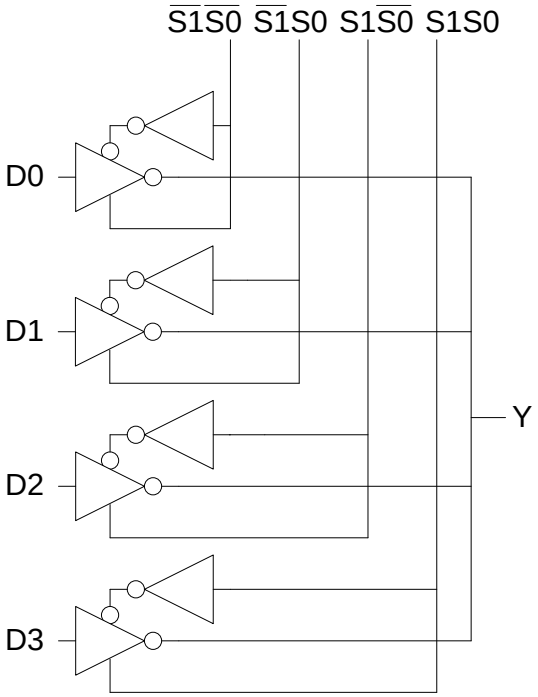




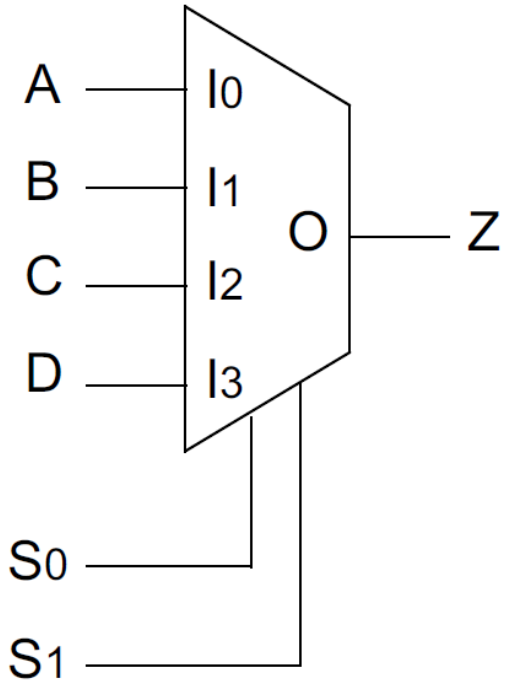
4:1 Multiplexer

- 4:1 mux chooses one of 4 inputs using two selects
  - Two levels of 2:1 muxes
  - Or four tristates





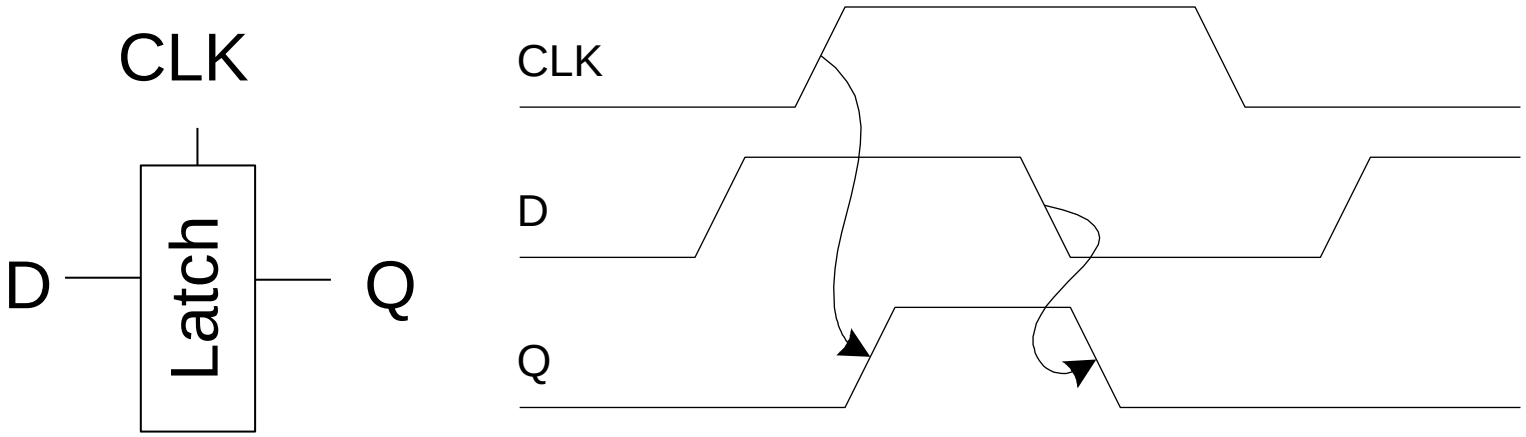
4:1 Multiplexer



S1	S0	Z
0	0	A
0	1	B
1	0	C
1	1	D

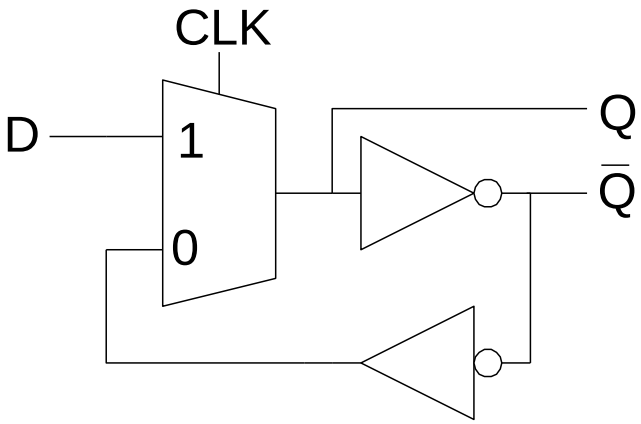
D Latch

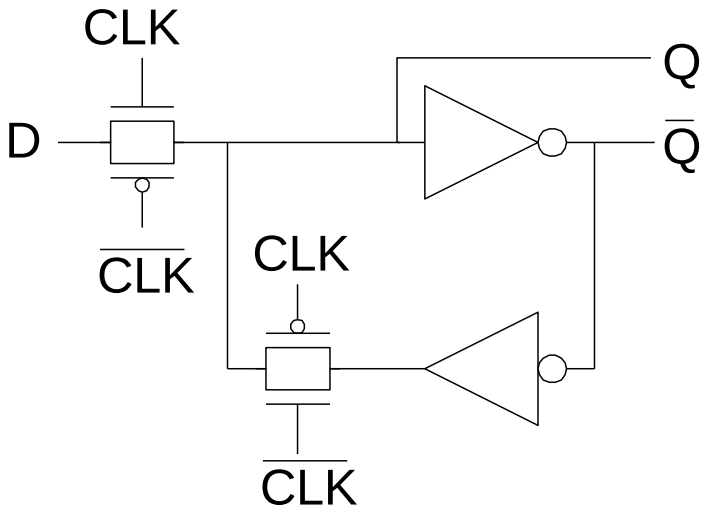
- When CLK = 1, latch is *transparent*
  - D flows through to Q like a buffer
- When CLK = 0, the latch is *opaque*
  - Q holds its old value independent of D
- a.k.a. *transparent latch* or *level-sensitive latch*



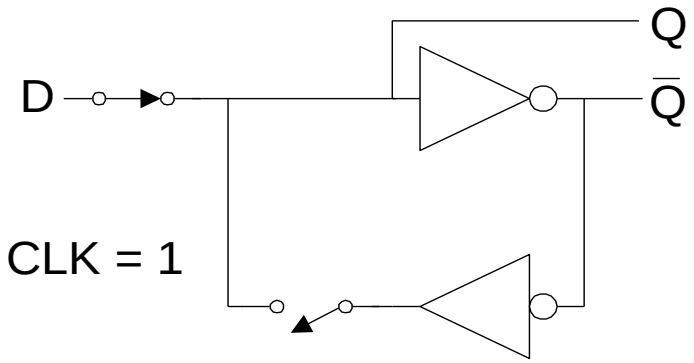
D Latch Design

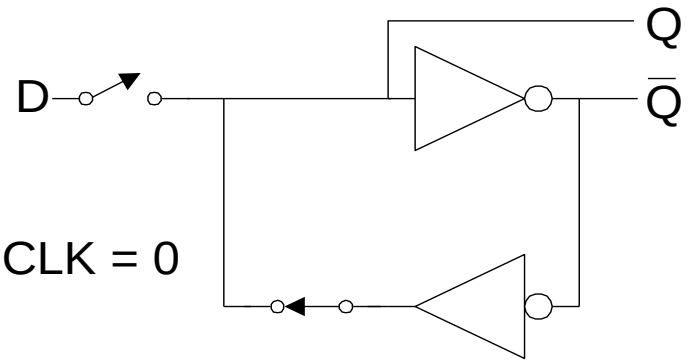
- Multiplexer chooses D or old Q

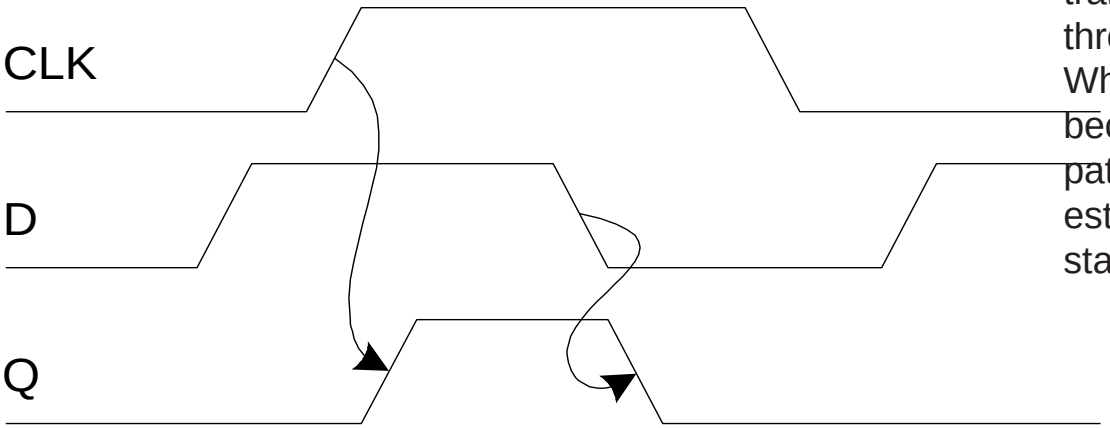




D Latch Operation



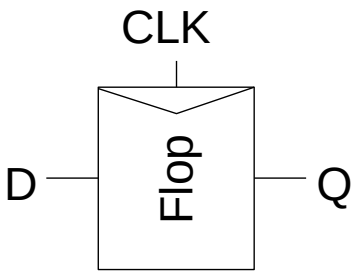


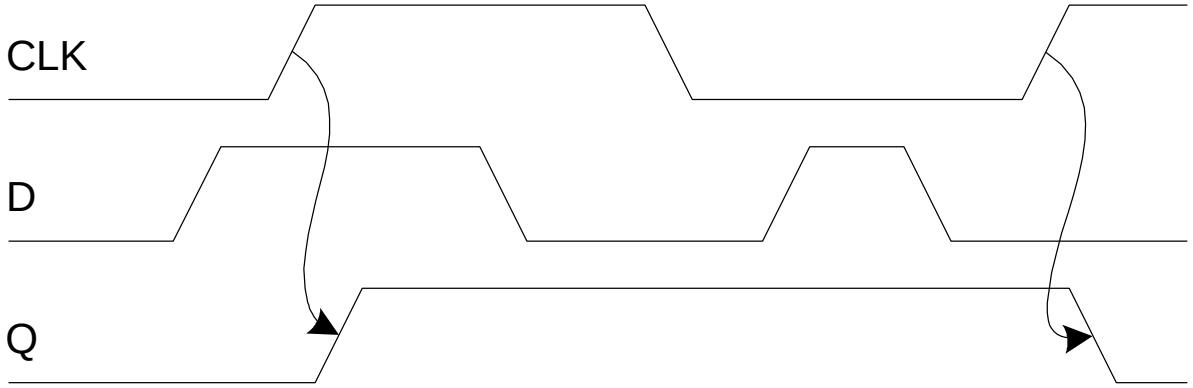


When *CLK* = 1, the latch is transparent and *D* flows through to *Q* (Figure 1.31(c)). When *CLK* falls to 0, the latch becomes opaque. A feedback path around the inverter pair is established to hold the current state of *Q* indefinitely

D Flip-flop

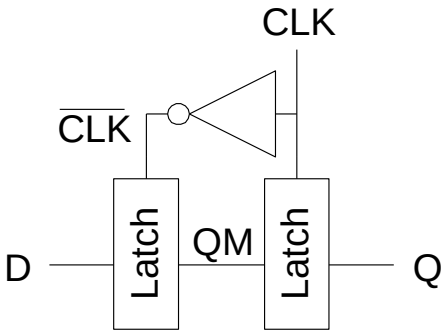
- When CLK rises, D is copied to Q
- At all other times, Q holds its value
- a.k.a. *positive edge-triggered flip-flop*, *master-slave flip-flop*

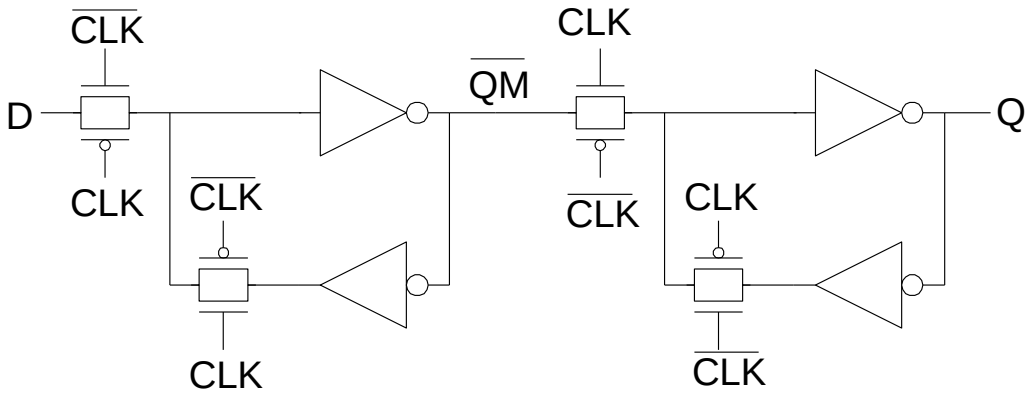




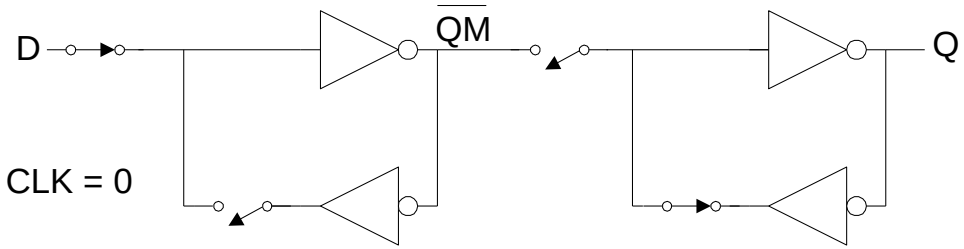
D Flip-flop Design

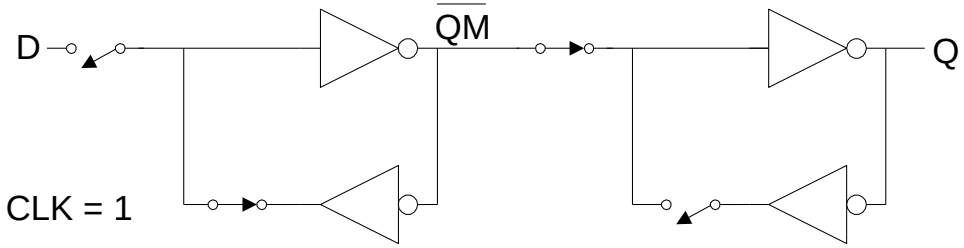
- Built from master and slave D latches

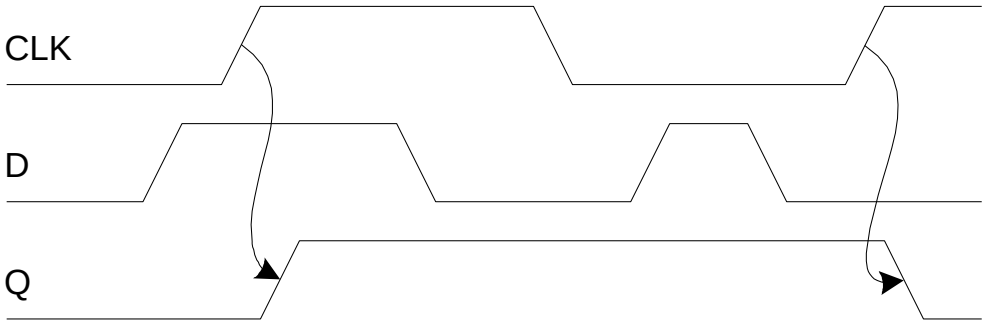




D Flip-flop Operation



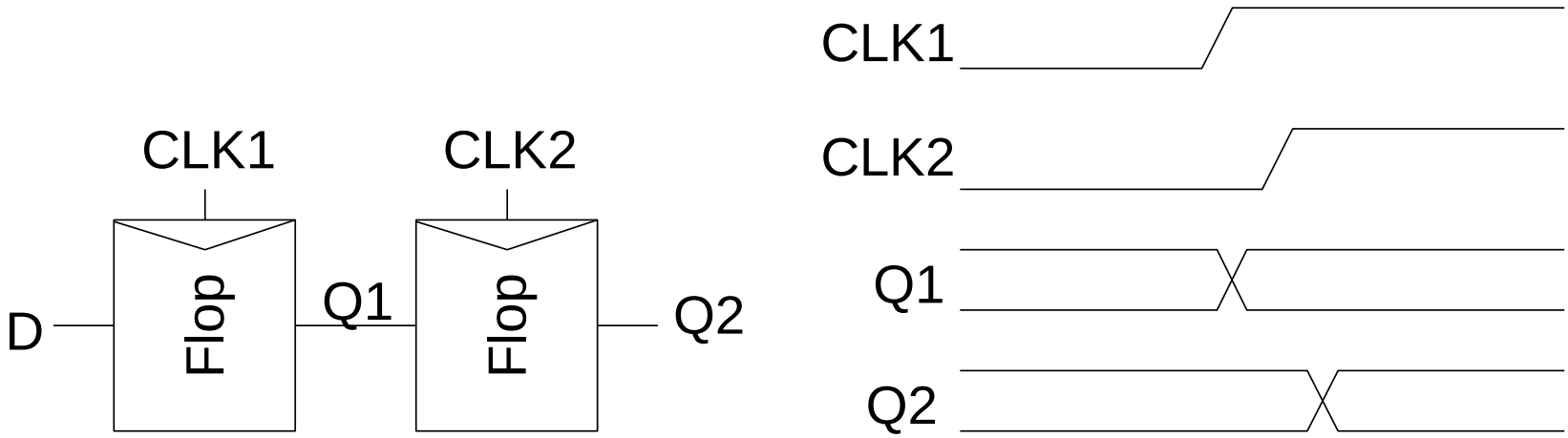




While *CLK* is low, the master negative-level-sensitive latch output ($\overline{QM}$) follows the *D* input while the slave positive-level-sensitive latch holds the previous value. When the clock transitions from 0 to 1, the master latch becomes opaque and holds the *D* value at the time of the clock transition. The slave latch becomes transparent, passing the stored master value ($\overline{QM}$) to the output of the slave latch (*Q*). The *D* input is blocked from affecting the output because the master is disconnected from the *D* input. When the clock transitions from 1 to 0, the slave latch holds its value and the master starts sampling the input again.

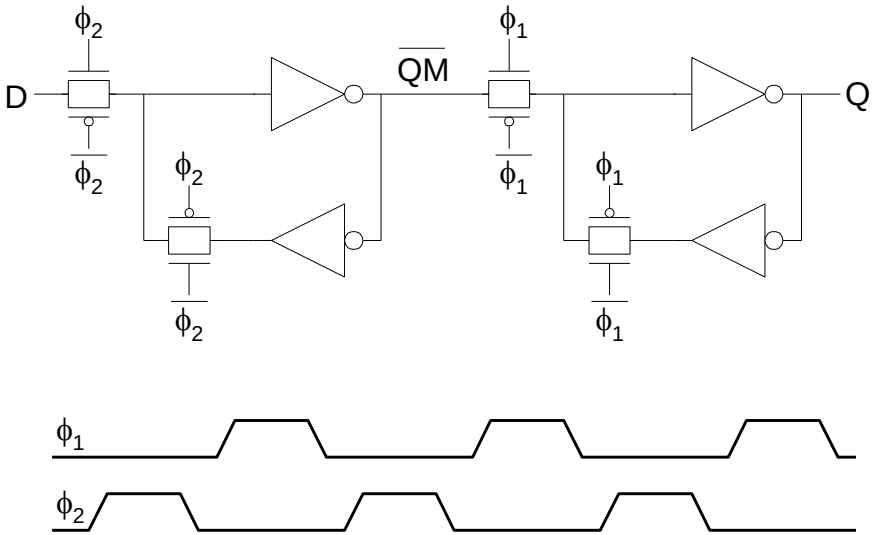
Race Condition

- Back-to-back flops can malfunction from clock skew
  - Second flip-flop fires late
  - Sees first flip-flop change and captures its result
  - Called *hold-time failure* or *race condition*



Nonoverlapping Clocks

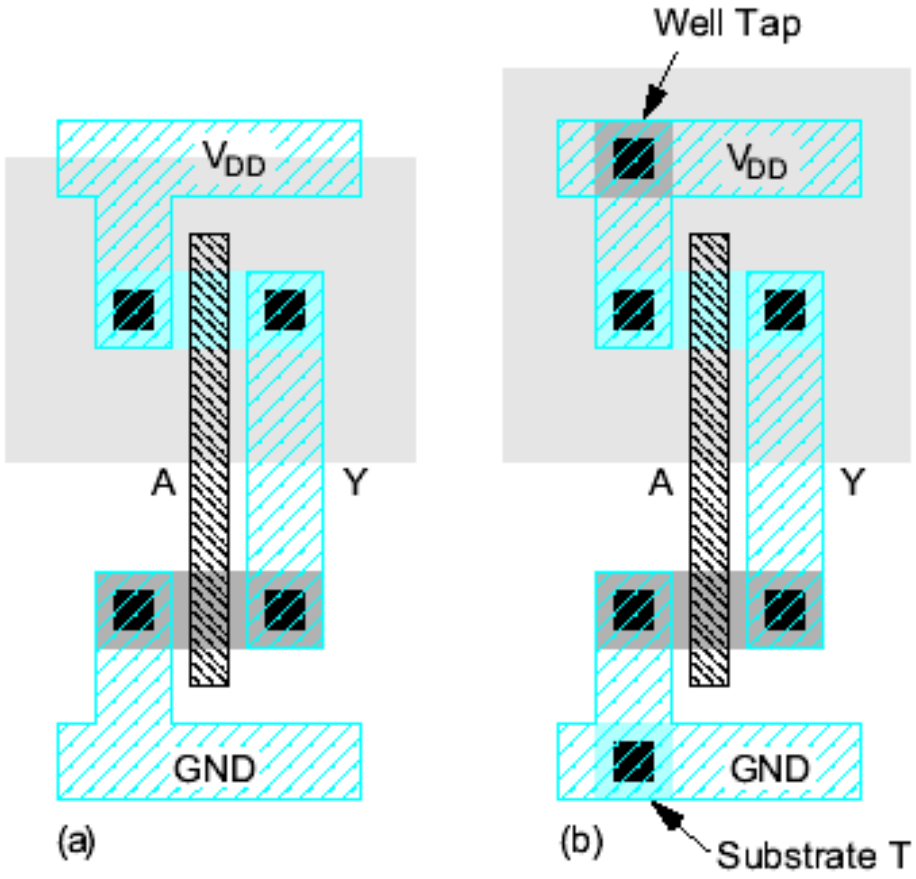
- Nonoverlapping clocks can prevent races
  - As long as nonoverlap exceeds clock skew
- We will use them in this class for safe design
  - Industry manages skew more carefully instead



Gate Layout

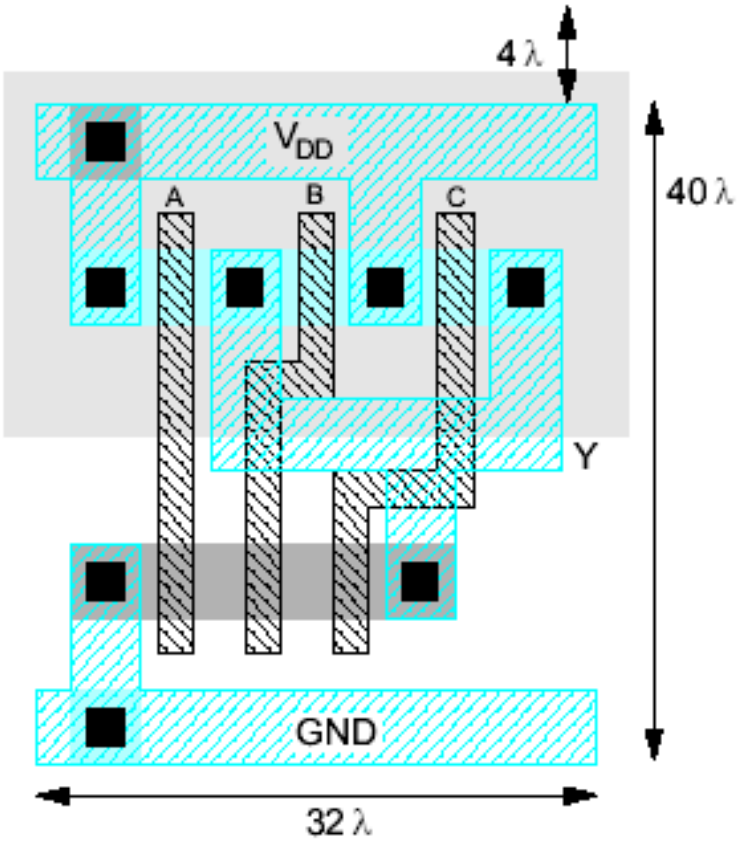
- Layout can be very time consuming
  - Design gates to fit together nicely
  - Build a library of standard cells
- Standard cell design methodology
  - V_{DD} and GND should abut (standard height)
  - Adjacent gates should satisfy design rules
  - nMOS at bottom and pMOS at top
  - All gates include well and substrate contacts

Example: Inverter



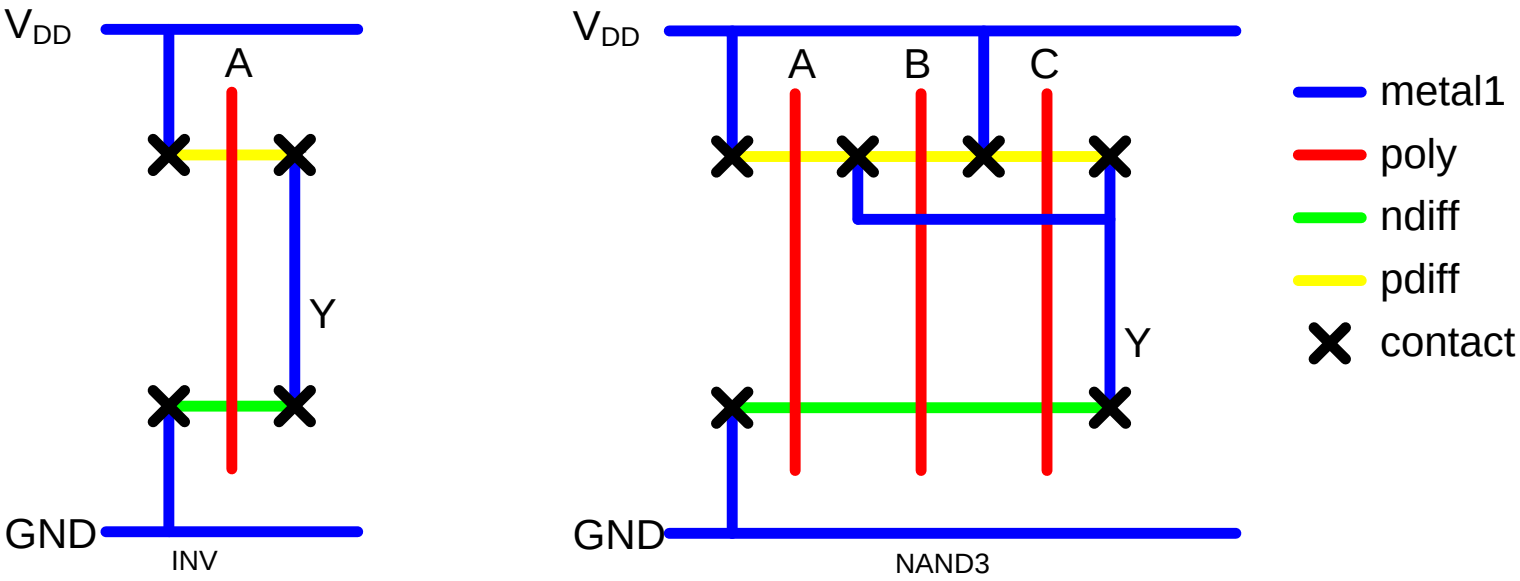
Example: NAND3

- Horizontal N-diffusion and p-diffusion strips
- Vertical polysilicon gates
- Metal1 V_{DD} rail at top
- Metal1 GND rail at bottom
- 32 λ by 40 λ



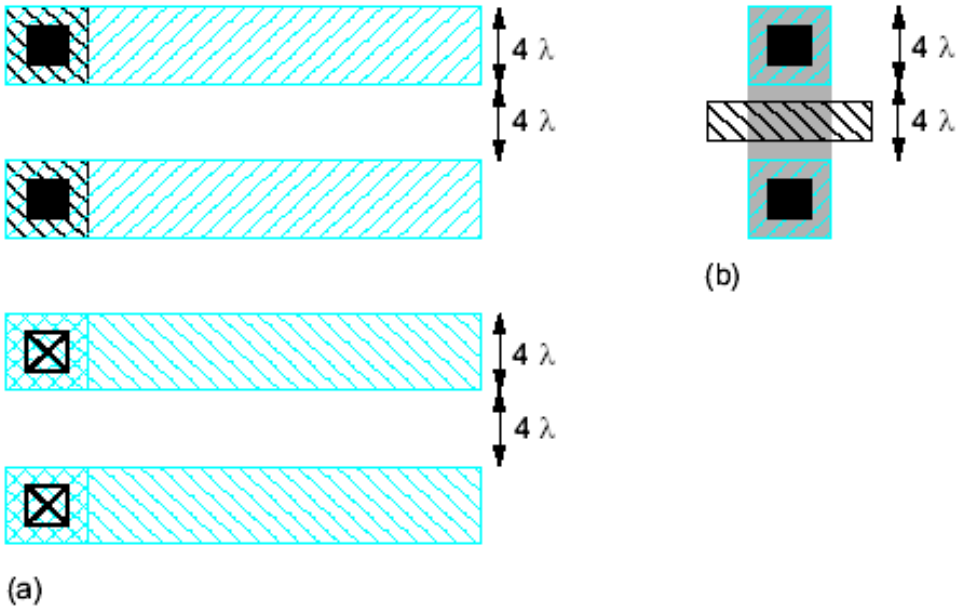
Stick Diagrams

- *Stick diagrams* help plan layout quickly
  - Need not be to scale
  - Draw with color pencils or dry-erase markers



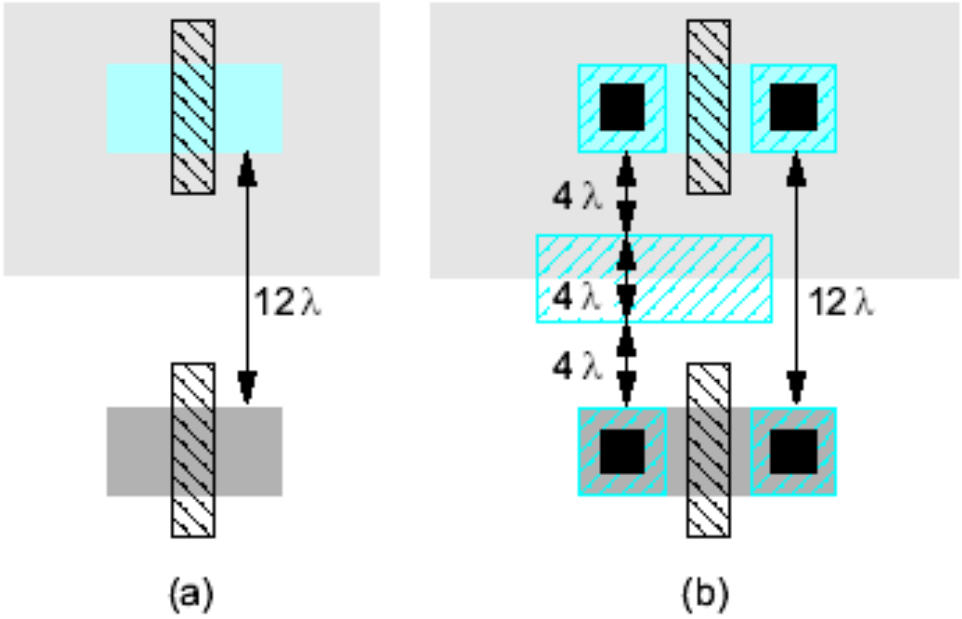
Wiring Tracks

- A *wiring track* is the space required for a wire
  - 4 λ width, 4 λ spacing from neighbor = 8 λ pitch
- Transistors also consume one wiring track



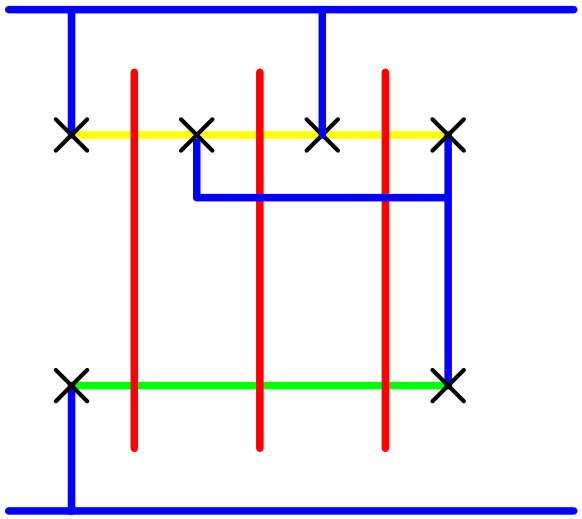
Well spacing

- Wells must surround transistors by 6 λ
  - Implies 12 λ between opposite transistor flavors
  - Leaves room for one wire track



Area Estimation

- Estimate area by counting wiring tracks
  - Multiply by 8 to express in λ



Example: O3AI

- Sketch a stick diagram for O3AI and estimate area
  - $Y = \overline{(A+B+C)\cdot D}$