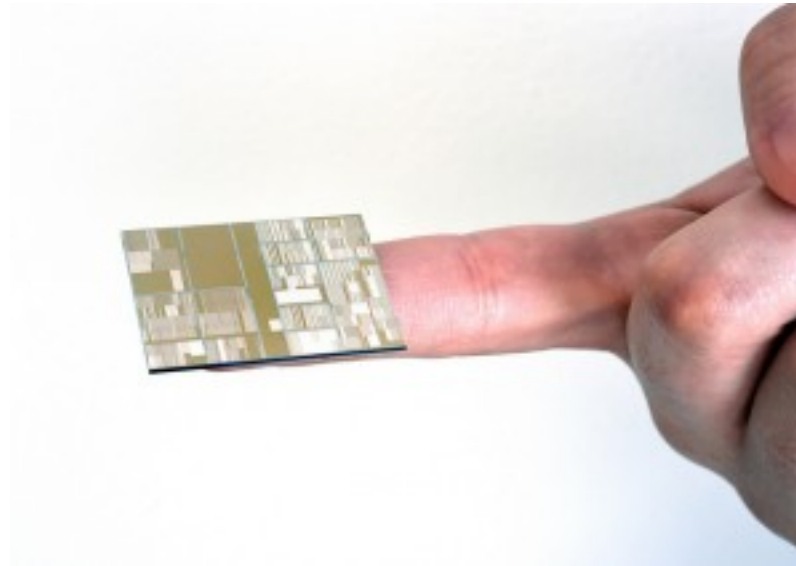


# Lecture 4

## - Integration -

### (1) CMOS Scaling

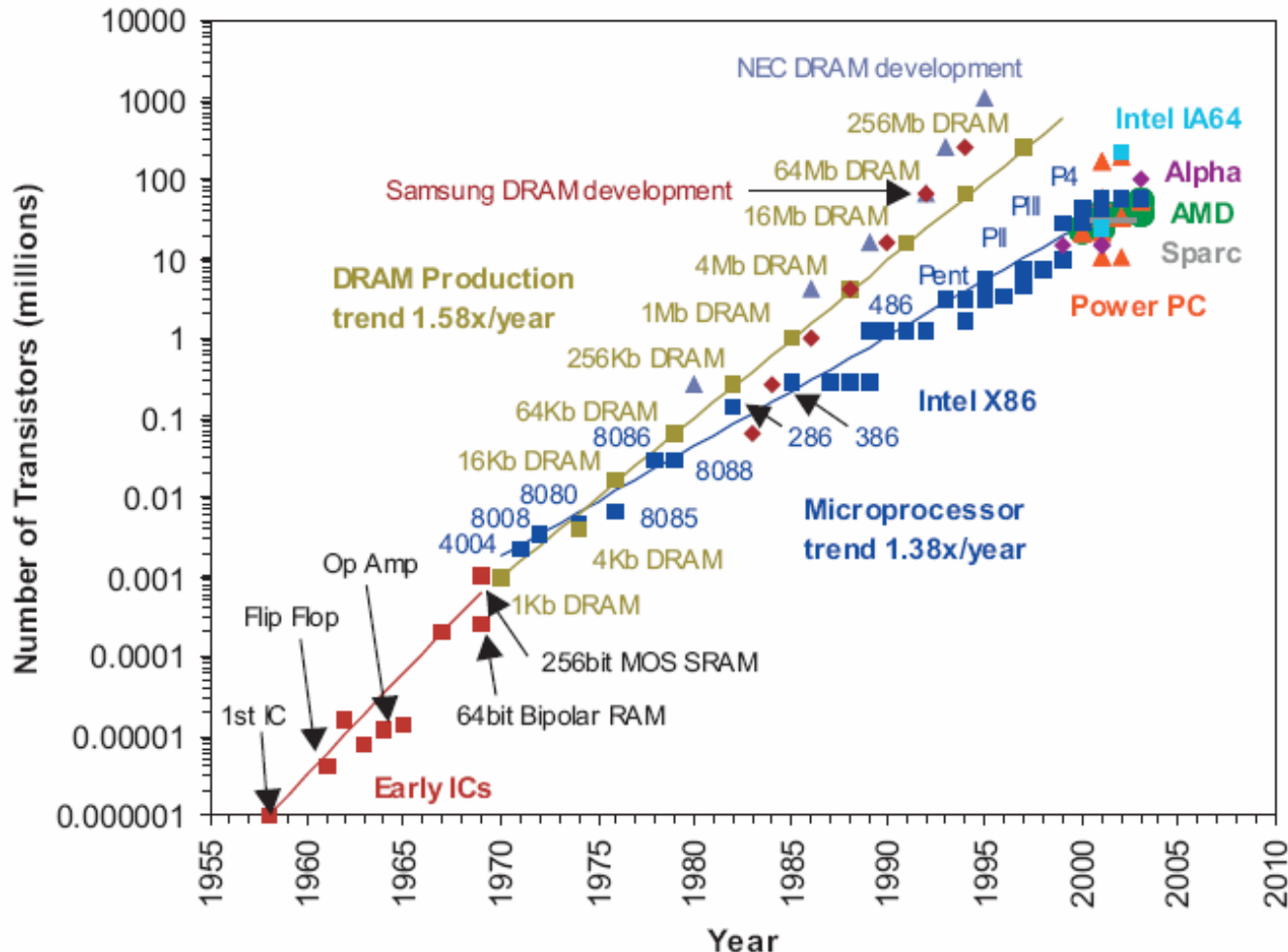


# Moore's Law

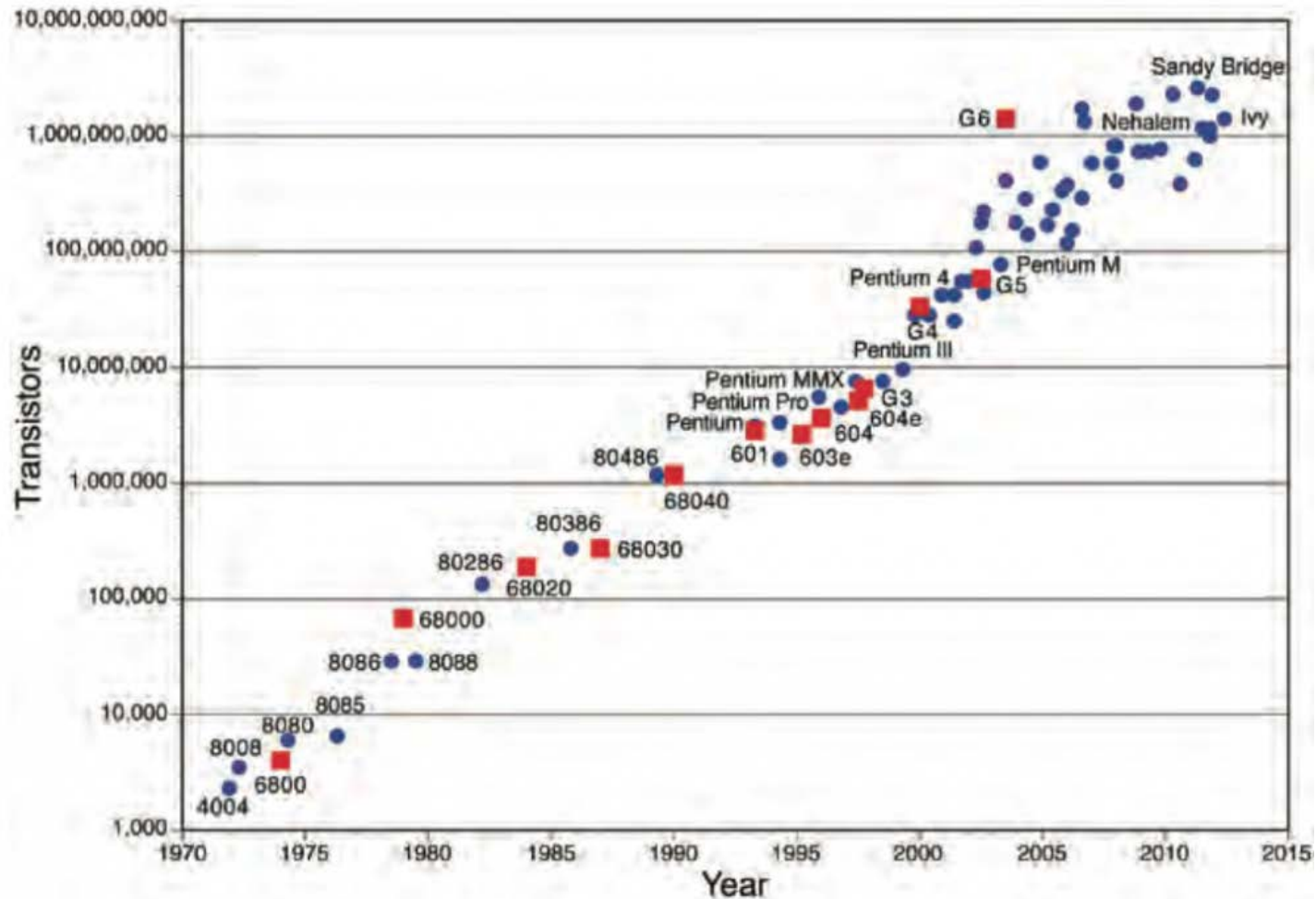
The number of transistors per square-inch doubles each 18 months



**Gordon Moore**  
Co-founder of Intel 1965



Moore, Gordon E. (1965).  
["Cramming more components onto integrated circuits"](#)  
[Electronics Magazine](#)

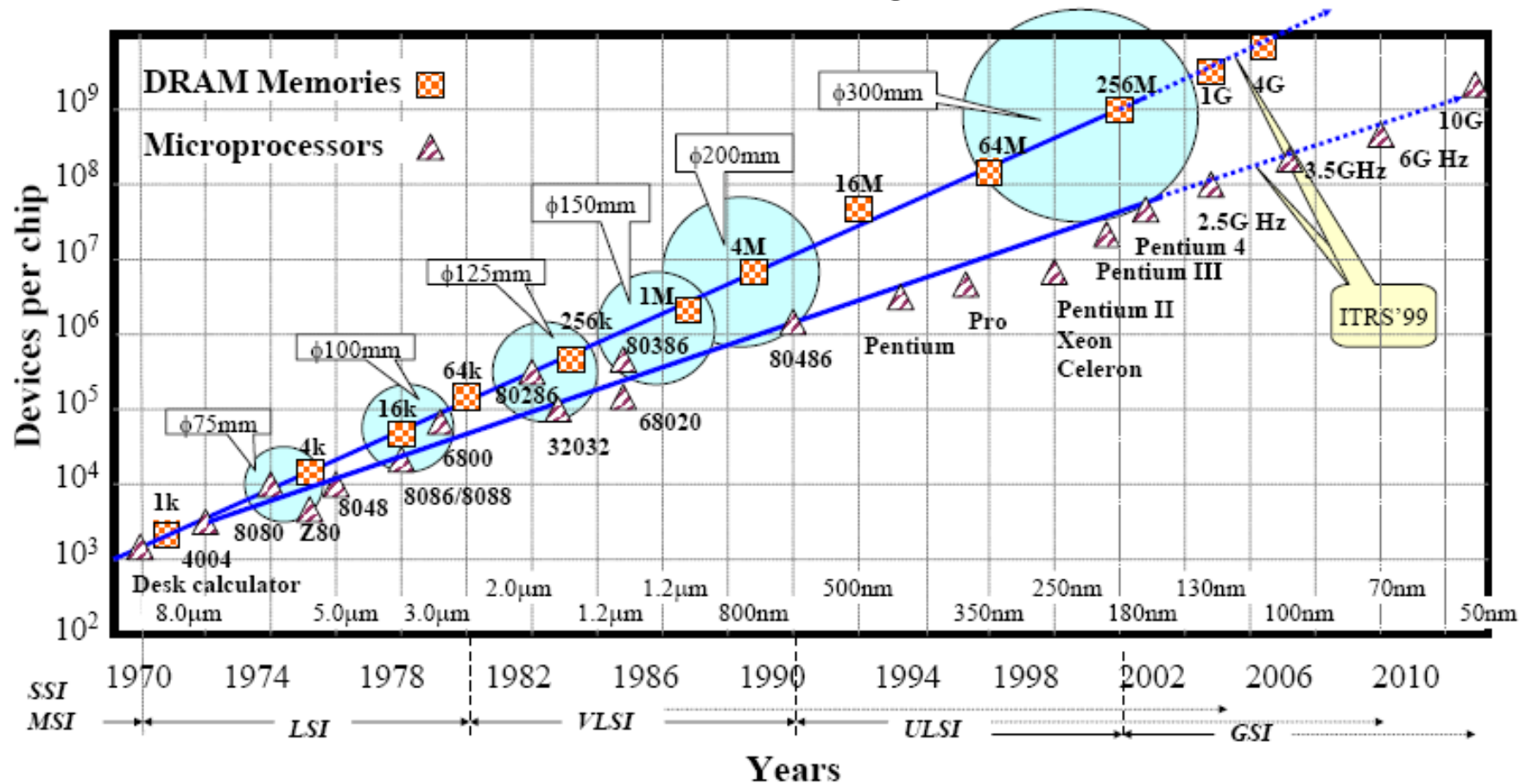


**FIGURE 1.** A plot of the increasing number of transistors per CPU confirms the accuracy of Moore's prediction. Note that the vertical axis is log scale.

Solid State Technology, Dec 2015

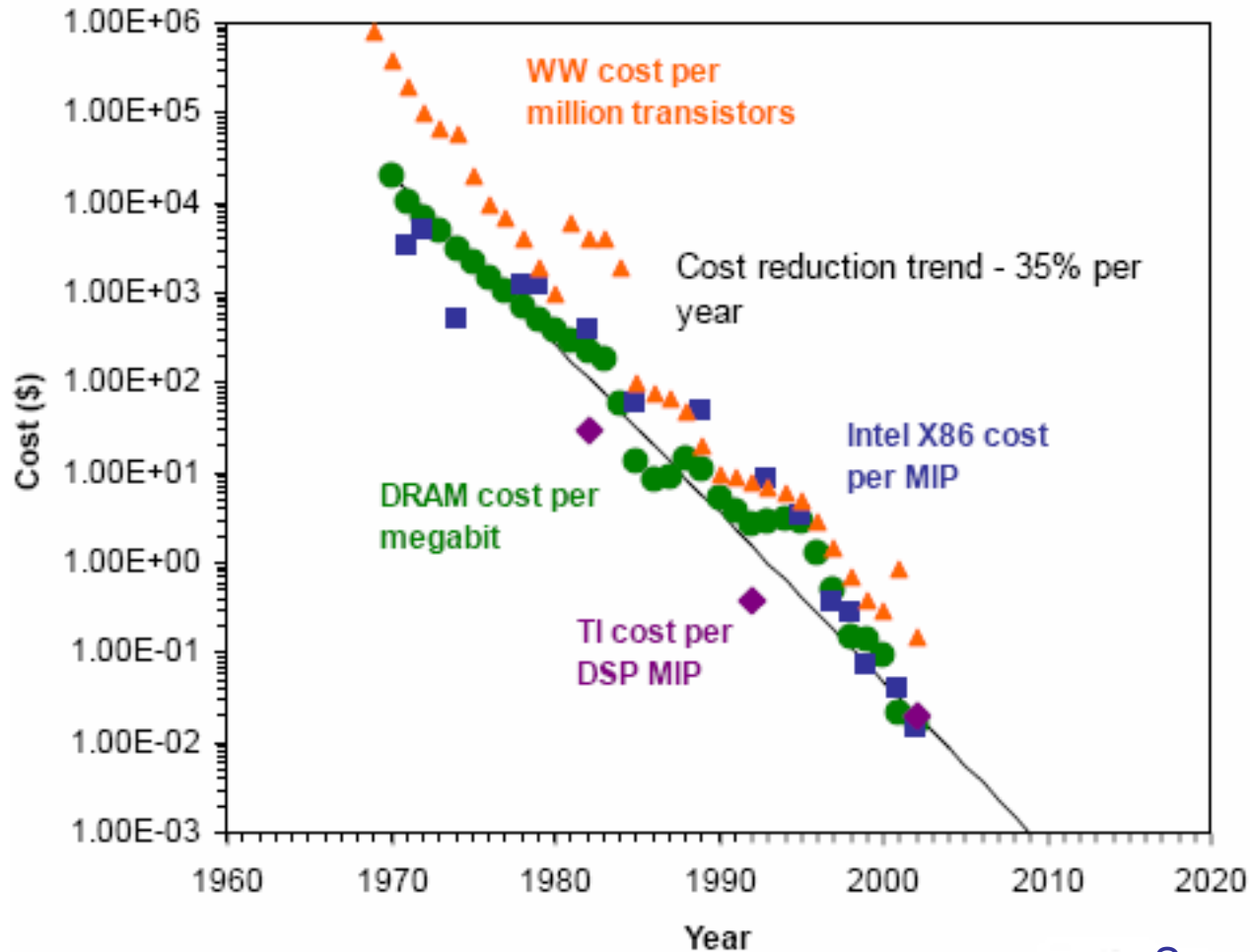
# IC Scaling

450mm?



Complexity	Number of Gates
Small-scale integration (SSI)	Fewer than 12
Medium-scale integration (MSI)	12 to 99
Large-scale integration (LSI)	100 to 9999
Very large-scale integration (VLSI)	10,000 to 99,999
Ultra large-scale integration (ULSI)	100,000 or more

# Price for each transistor keeps falling down





Source: Intel

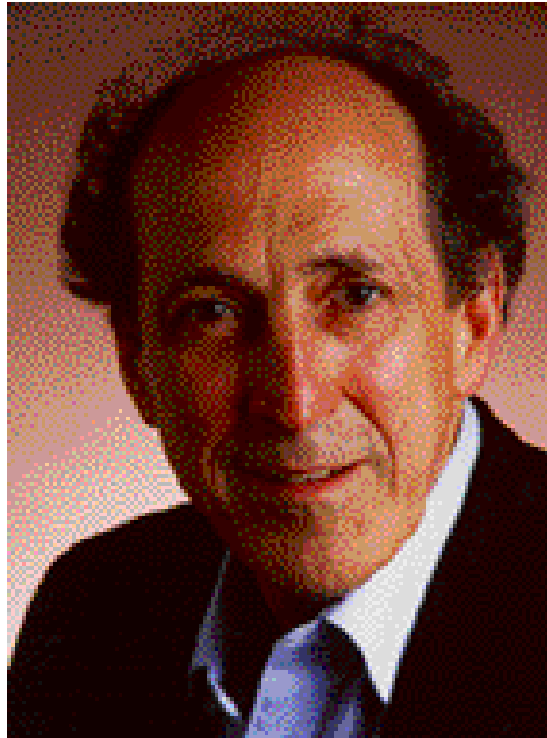
# If car industry follows Moore's Law

- If the car industry followed the Moore's Law in the past forty years.

Nowadays the car should be

- Price  12 cents/per car
- Speed  40,000 km/per hour
- Gas mileage  1200 km/per litter
- Capacity  400,000 person/per car

Source : Prof. T. P. Ma / Lo-SVP



**Robert Dennard, IBM**

1974 “Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions,” – Scaling Theory



The term "Moore's law" was coined around 1970 by the [Caltech](#) professor, [VLSI](#) pioneer, and [entrepreneur Carver Mead](#).

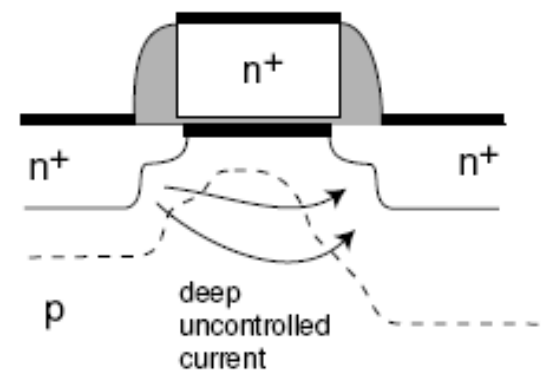
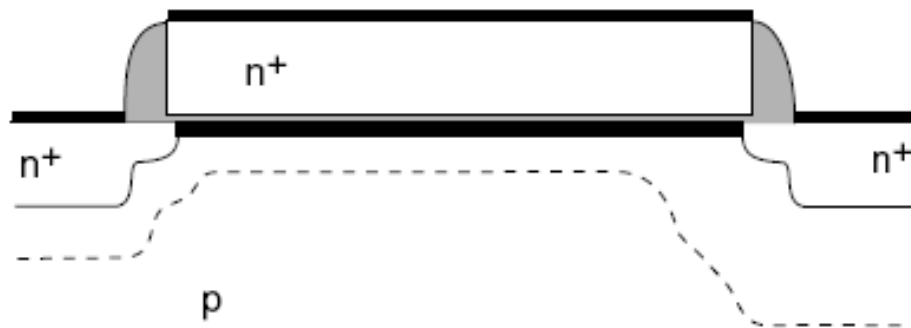
# MOSFET scaling

Several driving forces for scaling down size of MOSFET:

- higher density circuits: SSI, MSI, LSI, VLSI, ULSI, RLSI, ...
- higher performance:  $L \downarrow \Rightarrow I_D \uparrow \Rightarrow \tau_{switch} \downarrow$
- lower power consumption:  $L \downarrow \Rightarrow V_{DD} \downarrow$

Simple  $L$  scaling compromises *electrostatic integrity* and produces *punchthrough* (extreme case of short-channel effects):





To avoid punchthrough:

- $N_A \uparrow \Rightarrow V_T \uparrow \Rightarrow I_D \downarrow$
- $V_{DD} \downarrow \Rightarrow I_D \downarrow$
- $x_{ox} \downarrow \Rightarrow V_T \downarrow \Rightarrow I_D \uparrow$

Need smart way of scaling:

- constant field scaling
- constant voltage scaling
- generalized scaling

# Constant field scaling

Scale keeping vertical and horizontal electric fields constant.

Define: *scaling factor*  $S > 1$

parameter	scaling factor
device dimensions ( $L, W, x_{ox}$ )	$1/S$
doping level ( $N_A$ )	$S$
supply voltage ( $V_{DD}$ )	$1/S$

Consequences (use simple long-channel theory):

- gate capacitance:

$$C'_{gs} = C'_{ox} L' W' = S C_{ox} \frac{L}{S} \frac{W}{S} = \frac{C_{gs}}{S} \downarrow$$

- threshold voltage:

$$V_T' = V_{FB} + \phi_{sth} + \gamma\sqrt{\phi_{sth}} \simeq \frac{1}{C_{ox}'}\sqrt{2\epsilon_s q N_A' \phi_{sth}} \sim \frac{V_T}{\sqrt{S}} \downarrow$$

- drive current:

$$I_D' = \frac{W'}{2L'}\mu_e C_{ox}'(V_{DD}' - V_T')^2 = \frac{\frac{W}{S}}{2\frac{L}{S}}\mu_e S C_{ox}'\left(\frac{V_{DD}}{S} - \frac{V_T}{\sqrt{S}}\right)^2 = \frac{I_D}{S} \downarrow$$

- gate delay:

$$\tau' = \frac{C_{gs}' V_{DD}'}{I_D'} = \frac{\frac{C_{gs}}{S} \frac{V_{DD}}{S}}{\frac{I_D}{S}} = \frac{\tau}{S} \downarrow$$

- power-delay product or *switching energy*:

$$C'_{gs} V_{DD}'^2 = \frac{C_{gs}}{S} \left( \frac{V_{DD}}{S} \right)^2 = \frac{C_{gs} V_{DD}^2}{S^3} \downarrow\downarrow\downarrow$$

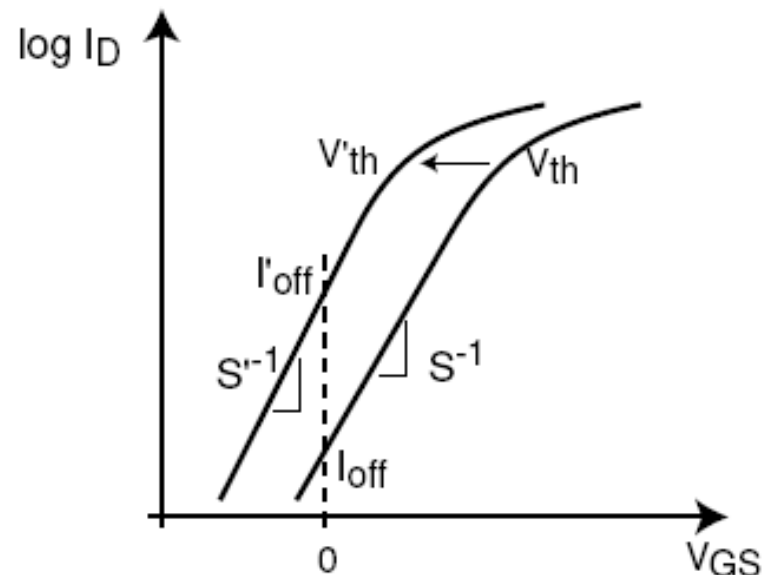
- switching energy density:

$$\frac{C'_{gs} V_{DD}'^2}{L'W'} = \frac{\frac{C_{gs} V_{DD}^2}{S^3}}{\frac{L}{S} \frac{W}{S}} = \frac{1}{S} \frac{C_{gs} V_{DD}^2}{LW} \downarrow$$

- inverse subthreshold slope:

$$n' = 1 + \frac{C'_{sth}}{C'_{ox}} = 1 + \frac{\sqrt{SC_{sth}}}{SC_{ox}} = 1 + \frac{C_{sth}}{\sqrt{SC_{ox}}} \downarrow$$

but since  $V_T \downarrow$ ,  $I_{off} \uparrow\uparrow$ .



Two key problems with constant field scaling:

- system designers don't want to scale  $V_{DD}$
- $I_{off} \uparrow \uparrow \Rightarrow$  more static power

# Constant voltage scaling

Scale all device dimensions but do not scale  $V_{DD}$ .

parameter	scaling factor
device dimensions ( $L, W, x_{ox}$ )	$1/S$
doping level ( $N_A$ )	$S$
supply voltage ( $V_{DD}$ )	1

Consequences (using long-channel theory):

figure of merit	scaling factor
$C_{gs}$	$1/S$
$V_{th}$	$1/\sqrt{S}$
$I_D$	$S$
$\tau$	$1/S^2$
$C_{gs}V_{DD}^2$	$1/S$
$C_{gs}V_{DD}^2/LW$	$S$

Features of constant voltage scaling:

- Performance  $\uparrow\uparrow$
- But:
  - It does not address  $I_{off}$  problem.
  - Electric field across oxide  $\uparrow$ :

$$\mathcal{E}_{ox} = \frac{V_{DD}}{x_{ox}} \propto S \uparrow$$

Reliability problems when  $\mathcal{E}_{ox} \simeq 4 \text{ MV/cm}$ .



- Electric field in semiconductor (at drain end of channel)  $\uparrow$ :

$$\mathcal{E}_m = \sqrt{\frac{(V_{DS} - V_{DSsat})^2}{l^2} + \mathcal{E}_{sat}^2} \propto S \uparrow$$

with

$$l^2 = \frac{\epsilon_s}{\epsilon_{ox}} x_{ox} x_j \propto S^{-2}$$

Reliability problems when  $\mathcal{E}_m \simeq 0.5 \text{ MV/cm}$ .

- Power density  $\uparrow \Rightarrow$  system power  $\uparrow$

# Generalized scaling

- scale oxide thickness more slowly than other device dimensions
- scale  $V_{DD}$  keeping  $\mathcal{E}_{ox}$  constant

parameter	scaling factor
$L, W$	$1/S$
$x_{ox}$	$1/R$
$N_A$	$S$
$V_{DD}$	$1/R$

with  $1 < R < S$ .

In generalized scaling:

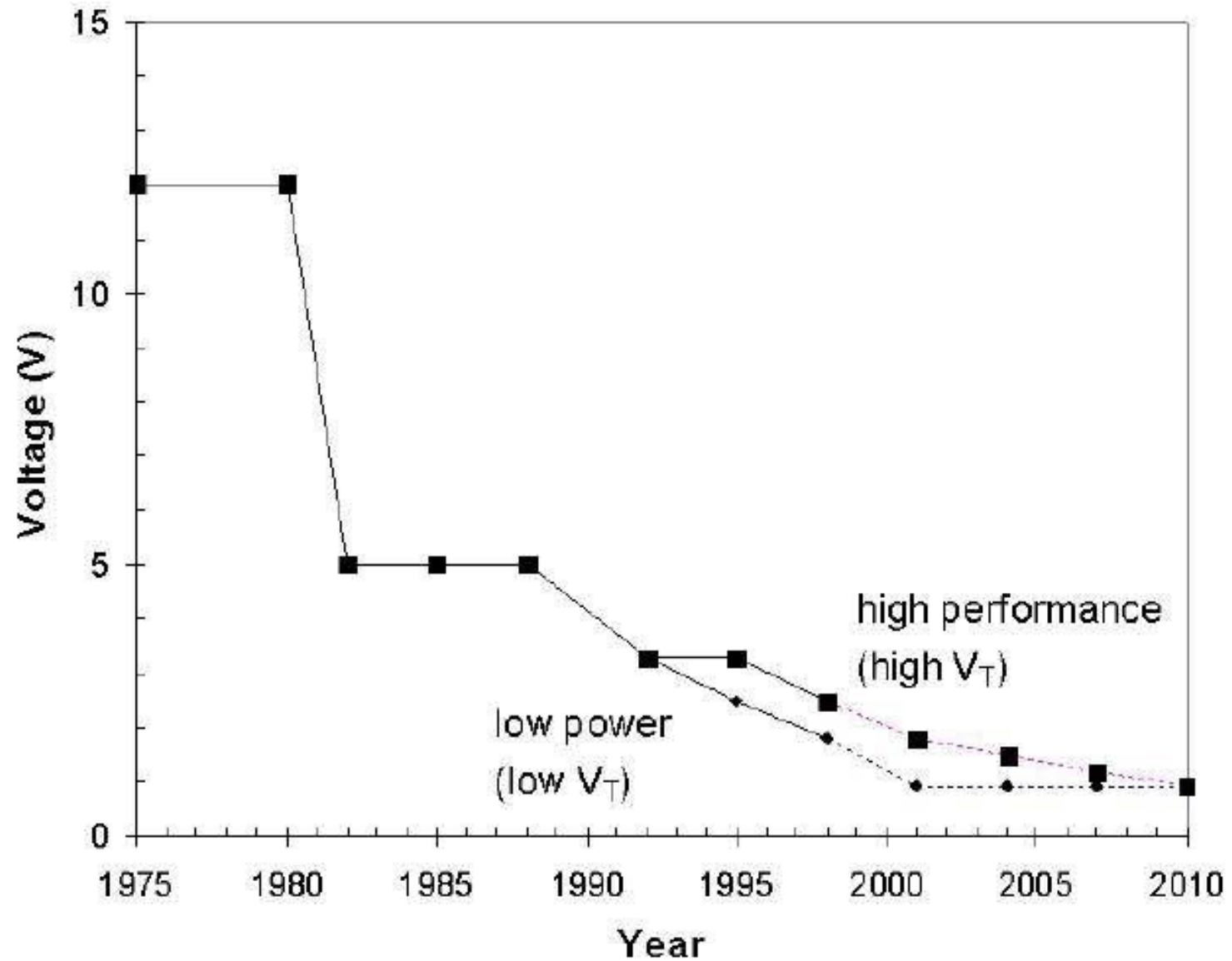
- $I_{off}$  problem alleviated by not scaling  $V_T$  so aggressively;  
*trade-off*: performance
- $V_{DD}$  scales;  
*trade-off*: performance

# Modern generalized scaling

- Concept of *generation*: every 2 years, new technology is deployed with 30% reduced transistor delay and twice as high transistor density (microprocessor performance doubling every 2 years).
- Everything scales:  $L$  ( $\downarrow$ ),  $W$  ( $\downarrow$ ),  $x_{ox}$  ( $\downarrow$ ),  $N_A$  ( $\uparrow$ ),  $x_j$  ( $\downarrow$ ), and  $V_{DD}$  ( $\downarrow$ ).
- Scaling goal: *extract maximum performance from each generation* (maximize  $I_{on}$ ), for a given amount of:
  - short-channel effects (DIBL), *and*
  - off-current

- Scaling goal: *extract maximum performance from each generation* (maximize  $I_{on}$ ), for a given amount of:
  - short-channel effects (DIBL), *and*
  - off-current
- Currently two technology flavors:
  - *high-performance*: high  $V_{DD}$  (high  $I_D$ , low  $\tau$ ), low  $V_T$  (high  $I_{off}$ );
  - *low-power*: low  $V_{DD}$  (low  $I_D$ , high  $\tau$ ), high  $V_T$  (low  $I_{off}$ ).

## Supply Voltage vs. Time



# Scaling

Scaling goal: *extract maximum performance from each generation* (maximize  $I_{on}$ ), for a given amount of:

- short-channel effects (DIBL), *and*
- off-current

To preserve *electrostatic integrity*, scaling has proceeded in a harmonious way:  $L$  ( $\downarrow$ ),  $W$  ( $\downarrow$ ),  $x_{ox}$  ( $\downarrow$ ),  $N_A$  ( $\uparrow$ ),  $x_j$  ( $\downarrow$ ), and  $V_{DD}$  ( $\downarrow$ ).

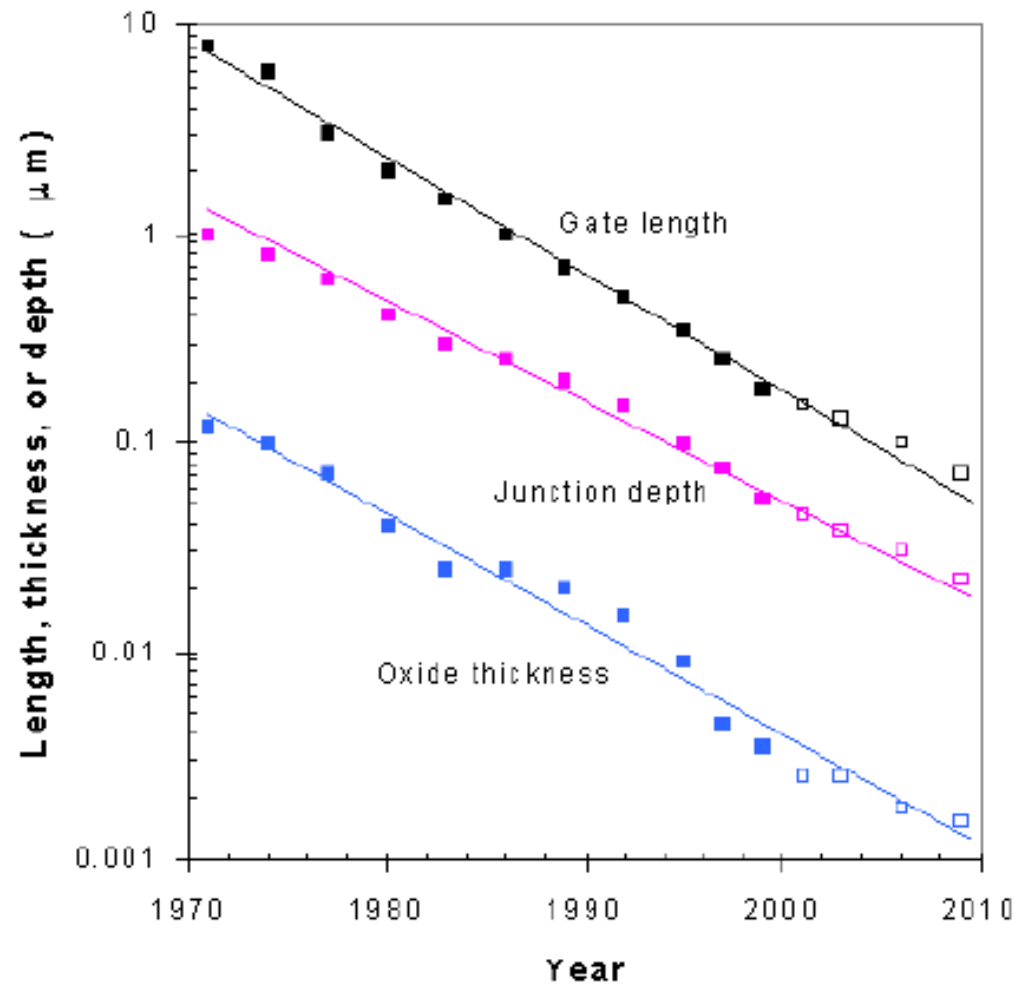
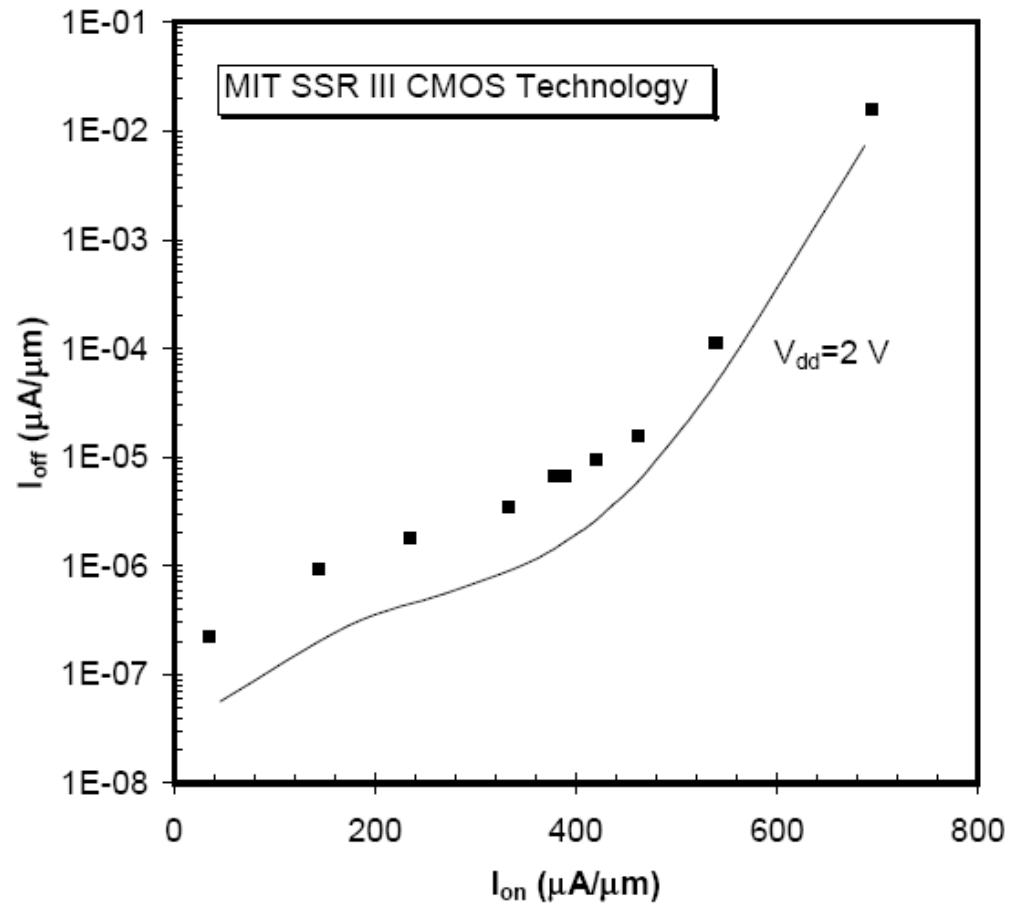


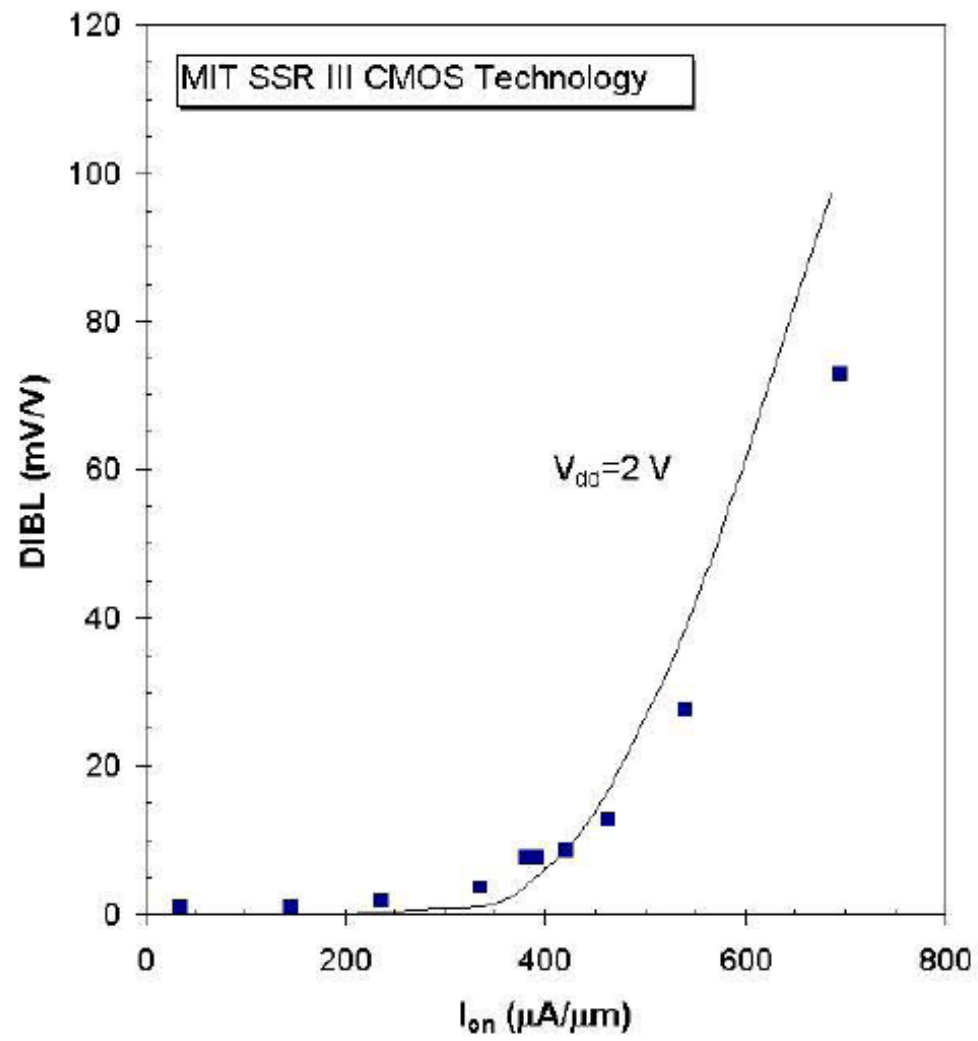


Illustration of key trade-offs:

- $I_{on}$  vs.  $I_{off}$



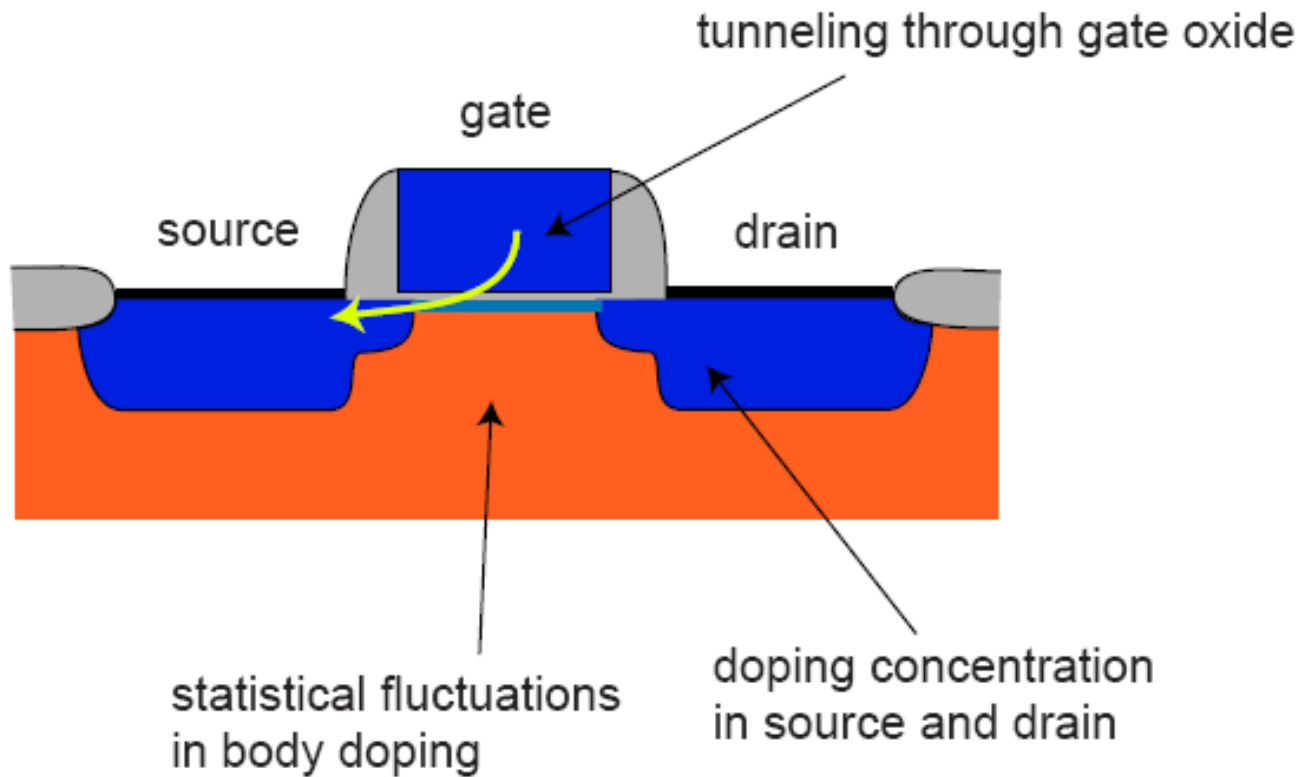
- $I_{on}$  vs.  $DIBL$



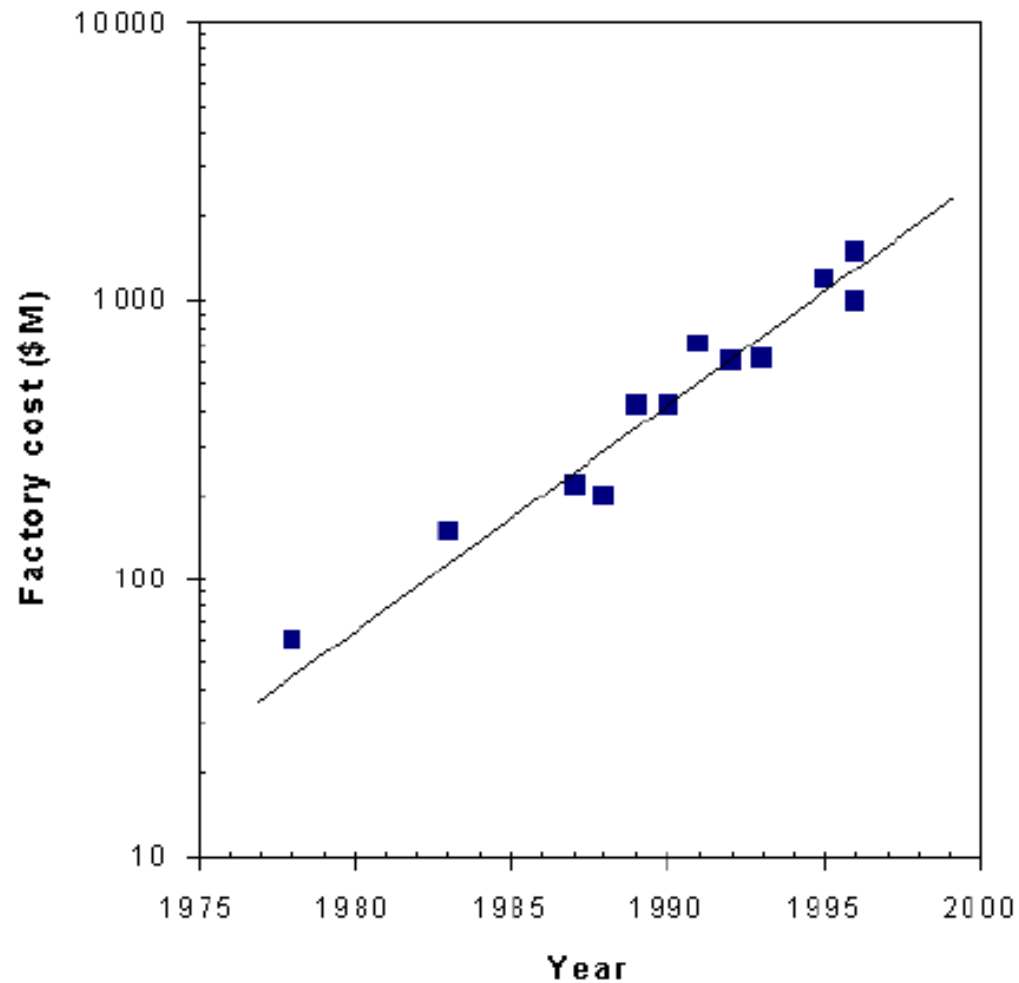
# Limits to scaling

Four kinds of limits:

- Thermodynamics: doping concentration in source and drain
- Physics: tunneling through gate oxide
- Statistics: statistical fluctuation of body doping
- Economics: factory cost



- Economics: factory cost also follows Moore's law!



- Physics: tunneling through gate oxide (most severe limit)

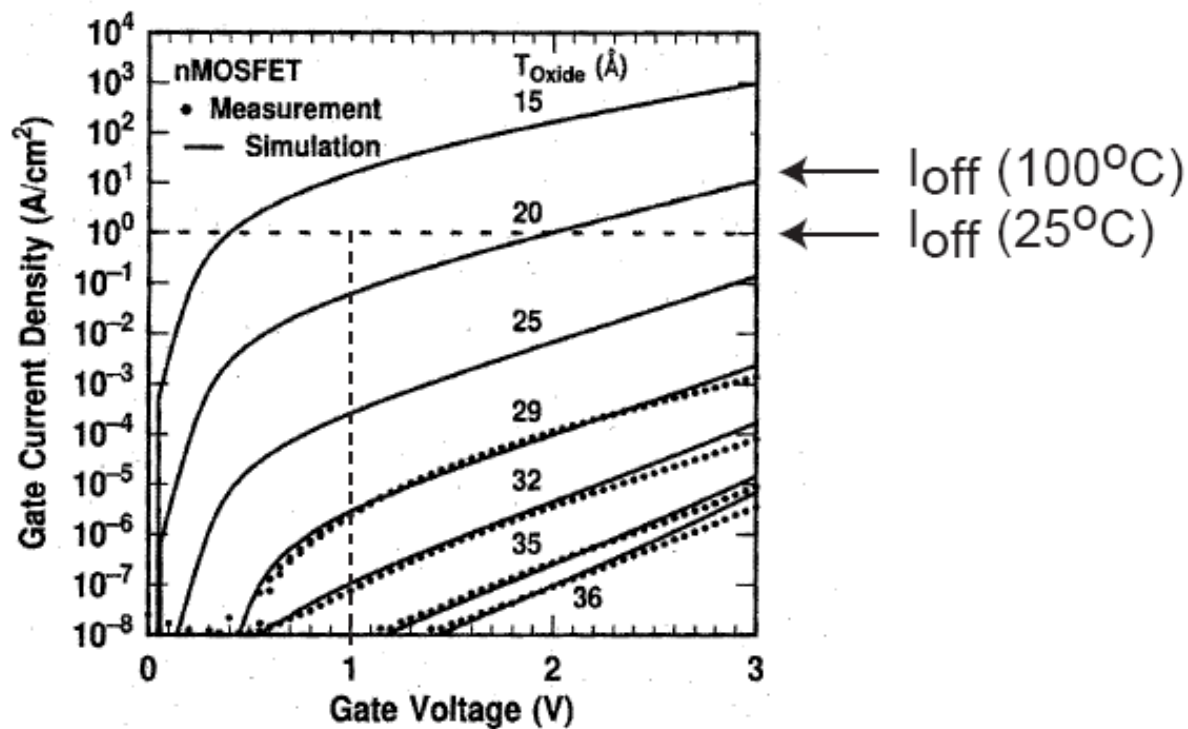


Fig. 2. Measured and simulated  $I_G$ - $V_G$  characteristics under inversion conditions of four nMOSFET's. The dotted line indicates the  $1 \text{ A}/\text{cm}^2$  limit for leakage current as discussed in the text.

Figure 13 on p. 491 in: Taur, Y., et al. "CMOS Scaling into the Nanometer Regime." *Proceedings of the IEEE* 85, no. 4 (1997): 486-504. © 1997 IEEE.

- Oxide's thickness limit when:

$$I_{gate} \simeq I_{off} \text{ @ } V_{DD} \simeq 1 \text{ V, } T_{oper}(\simeq 100^\circ C)$$

- Translates to limiting gate current:

$$I_{gate}(25^\circ C) \simeq 100 \text{ pA}$$

- Limiting gate current density:

$$A \simeq 0.1 \mu m \times 0.1 \mu m = 10^{-10} \text{ cm}^2 \Rightarrow J_{gate}(25^\circ C) \simeq 1 \text{ A/cm}^2$$

- Limiting  $x_{ox} \simeq 1.6 \text{ nm} \Rightarrow L \sim 35 - 50 \text{ nm}$
- Solution: *high-dielectric constant gate insulator*

□ To go beyond this, need:

- new materials that squeeze more performance out of existing device architecture
  - new channel materials: strained Si, Si/SiGe heterostructures
  - new gate insulators: high-K dielectric, such as HfO<sub>2</sub>
  - new gate conductors: metal gate, such as fully silicided gate
- new device architecture (SOI, double gate, trigate) to improve electrostatic integrity



# Evolution of MOSFET design

- PMOS with metal gate:

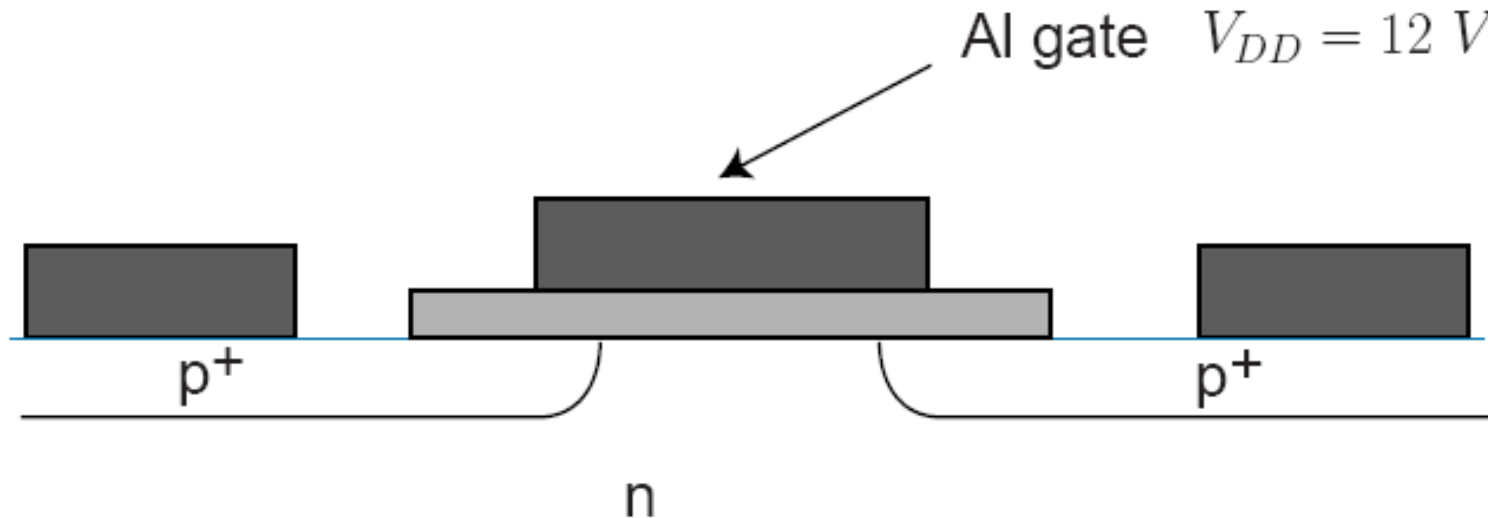
circa~early 70's

$L \sim 20 \mu m$

$x_{ox} \sim 1000 \text{ \AA}$

$x_j \sim 3 \mu m$

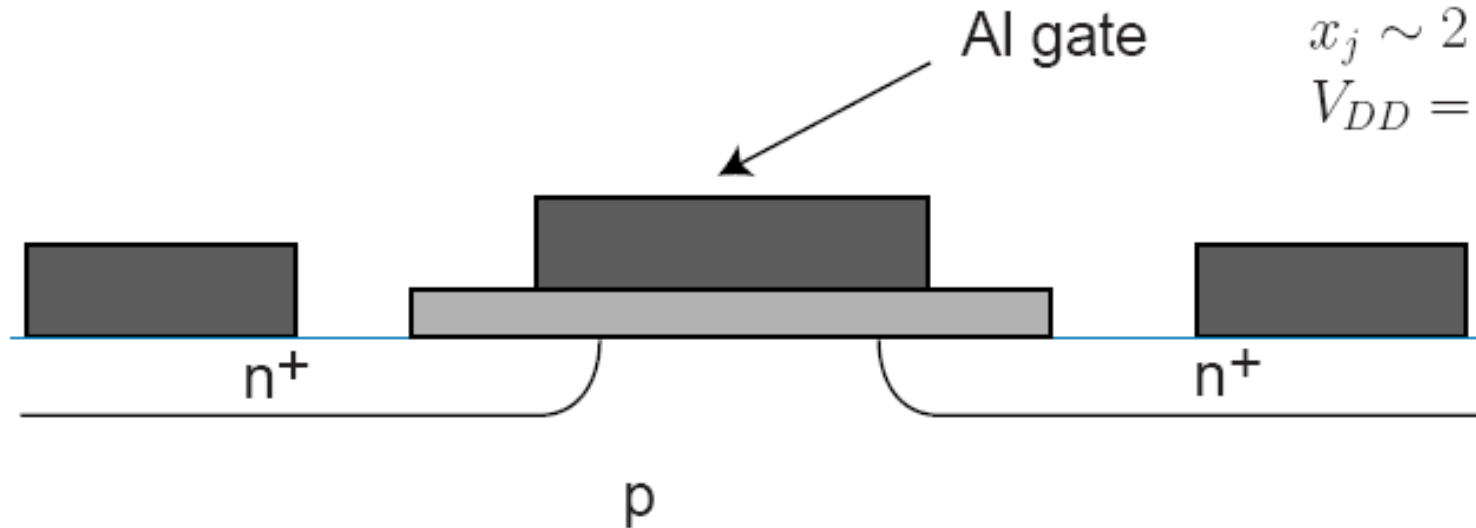
$V_{DD} = 12 V$



Main point:  $\text{Na}^+$  contamination made NMOS devices to have too negative a threshold voltage

- NMOS with metal gate:

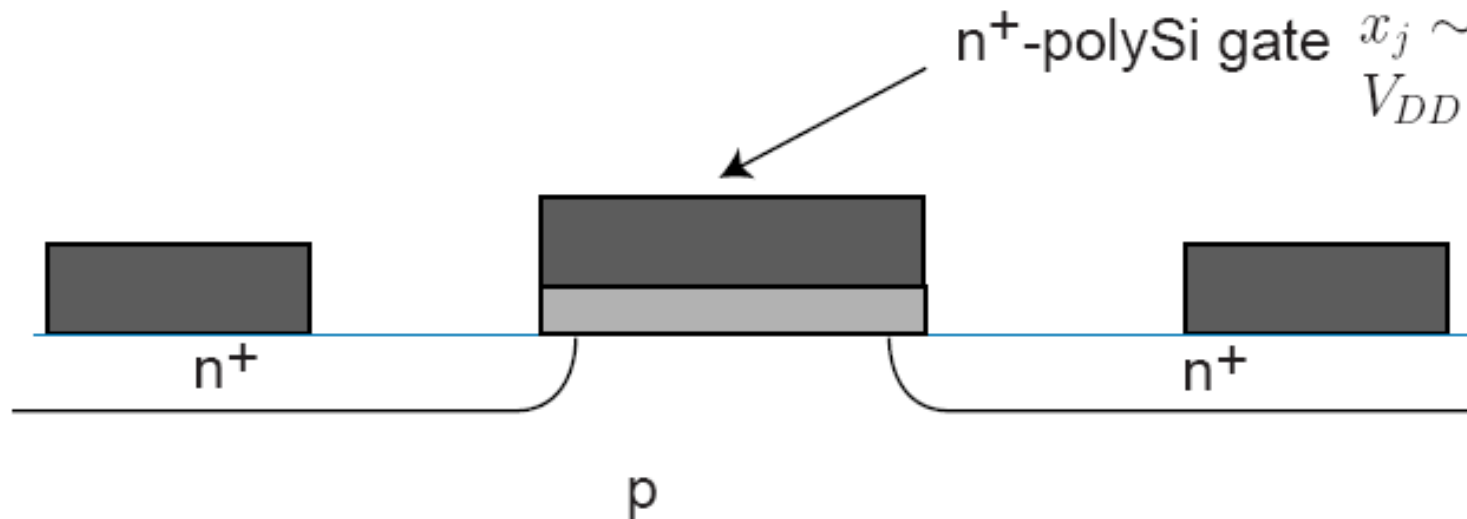
circa~1975  
 $L \sim 15 \mu m$   
 $x_{ox} \sim 600 \text{ \AA}$   
 $x_j \sim 2 \mu m$   
 $V_{DD} = 12 V$



Main point: with  $Na^+$  contamination under control, NMOS devices became possible (higher performance).

- CMOS with self-aligned polySi gate:

circa  $\sim 1980$   
 $L \sim 2 \mu m$   
 $x_{ox} \sim 400 \text{ \AA}$   
 $x_j \sim 1 \mu m$   
 $V_{DD} = 5 V$



Main point: self-aligned process allows tighter overlap between gate and  $n^+$  regions and results in lower parasitic capacitance.

- Lightly-doped drain MOSFET (LDD-MOSFET):

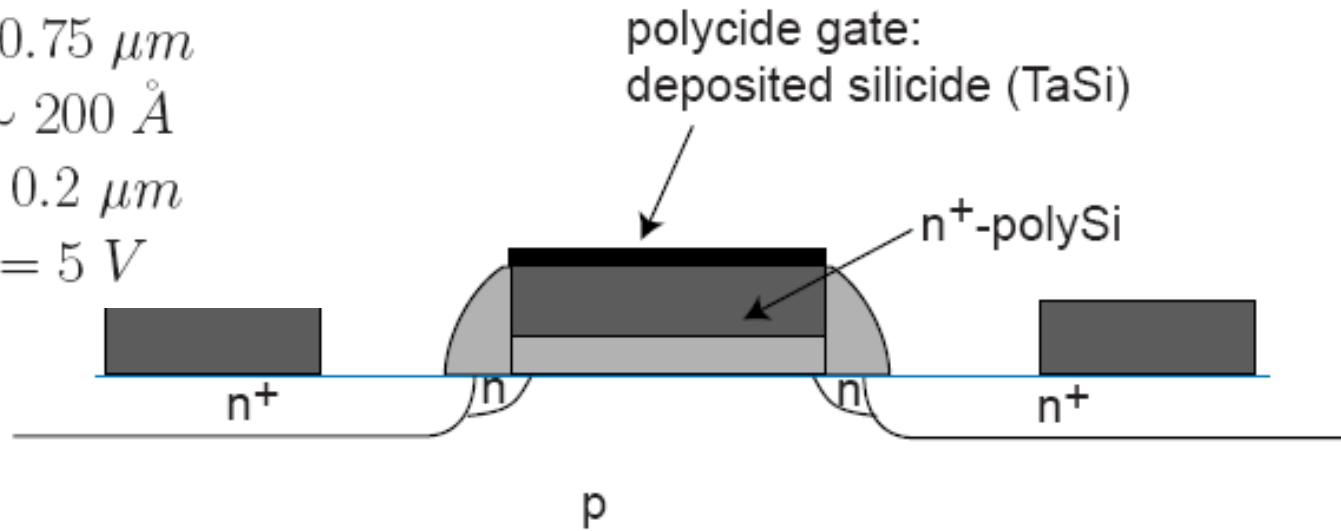
circa~1985

$L \sim 0.75 \mu m$

$x_{ox} \sim 200 \text{ \AA}$

$x_j \sim 0.2 \mu m$

$V_{DD} = 5 V$



Main point: lightly-doped n-region on drain side reduces electric field there and allows a high  $V_{DD}$  to be used.

- Salicide (self-aligned silicide) MOSFET:

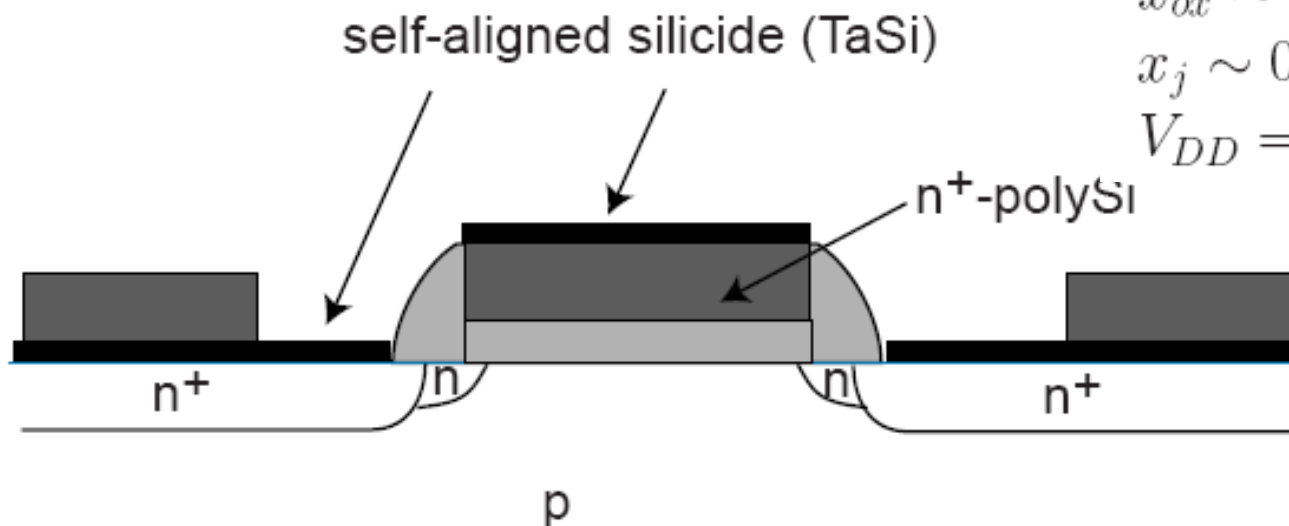
circa~1989

$L \sim 0.4 \mu m$

$x_{ox} \sim 125 \text{ \AA}$

$x_j \sim 0.15 \mu m$

$V_{DD} = 3.3 V$



Main point: salicided gate, source and drain reduces all parasitic resistances.

- MOSFET with p-pocket or halo implants:

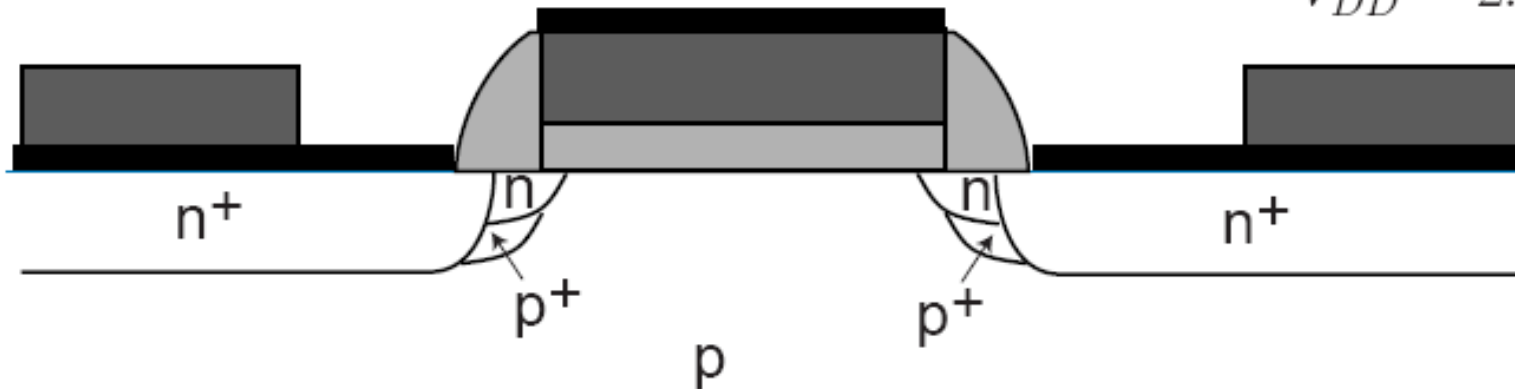
circa~1994

$L \sim 0.15 \mu m$

$x_{ox} \sim 60 \text{ \AA}$

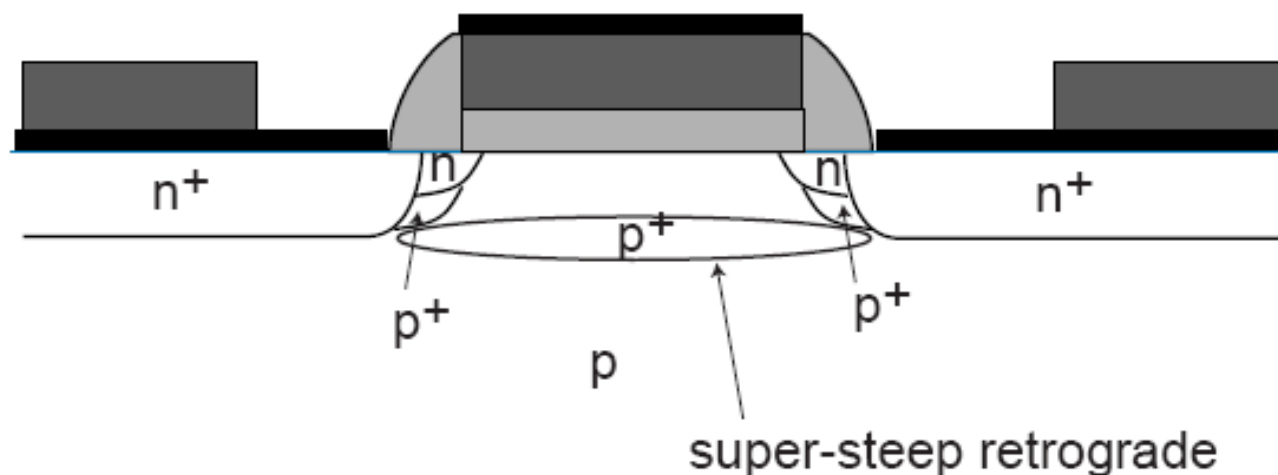
$x_j \sim 0.08 \mu m$

$V_{DD} = 2.5 V$



Main point:  $p^+$  pockets control short-channel effects.

- Sub-0.1  $\mu m$  MOSFET:



circa~late 90's (manufacturing in early 00's)

$$L < 0.1 \mu m$$

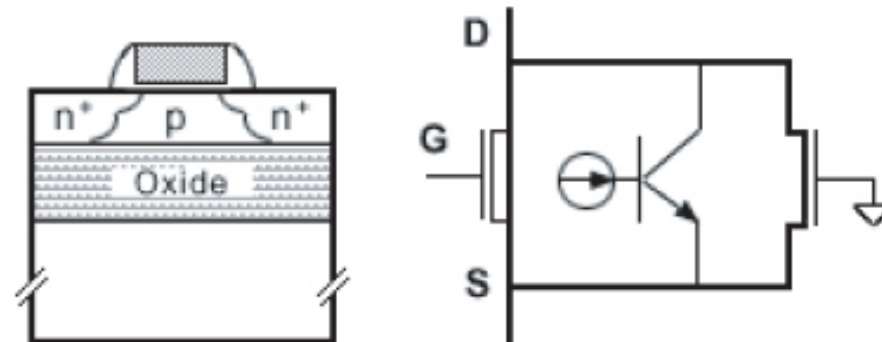
$$x_{ox} \sim 30 \text{ \AA}$$

$$x_j \sim 0.06 \mu m$$

$$V_{DD} = 0.8 - 1.5 V$$

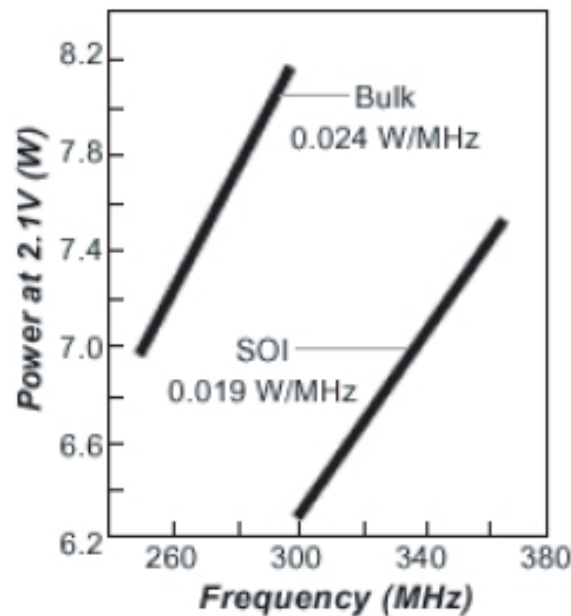
Main point: p<sup>+</sup>-super-steep retrograde body doping controls short-channel effects while preserving high mobility.

New device architecture: Silicon-on-Insulator (SOI)



*Schematic of nFET on SOI and equivalent devices.  
Adapted from Shahidi et al., **Proc. ISSCC**, 1999 (426).*



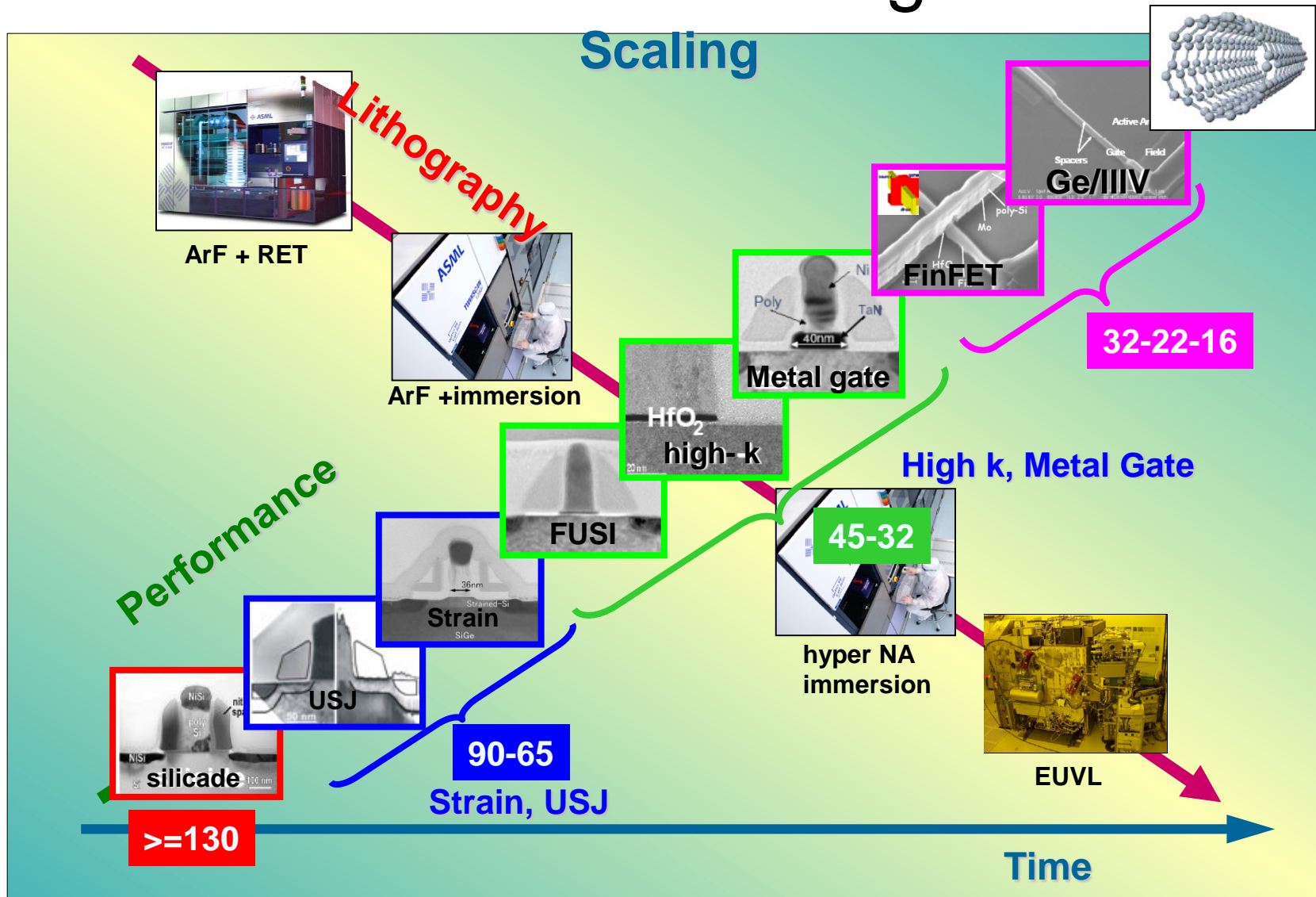


Power vs. Frequency  
Adapted from Shahidi et al., **Proc. ISSCC**, 1999 (426).

A number of issues associated with existence of buried oxide:

- reduced junction capacitance
- floating body: kink effect, extra drive ( $V_{BS} > 0$  during switching)
- increased thermal resistance

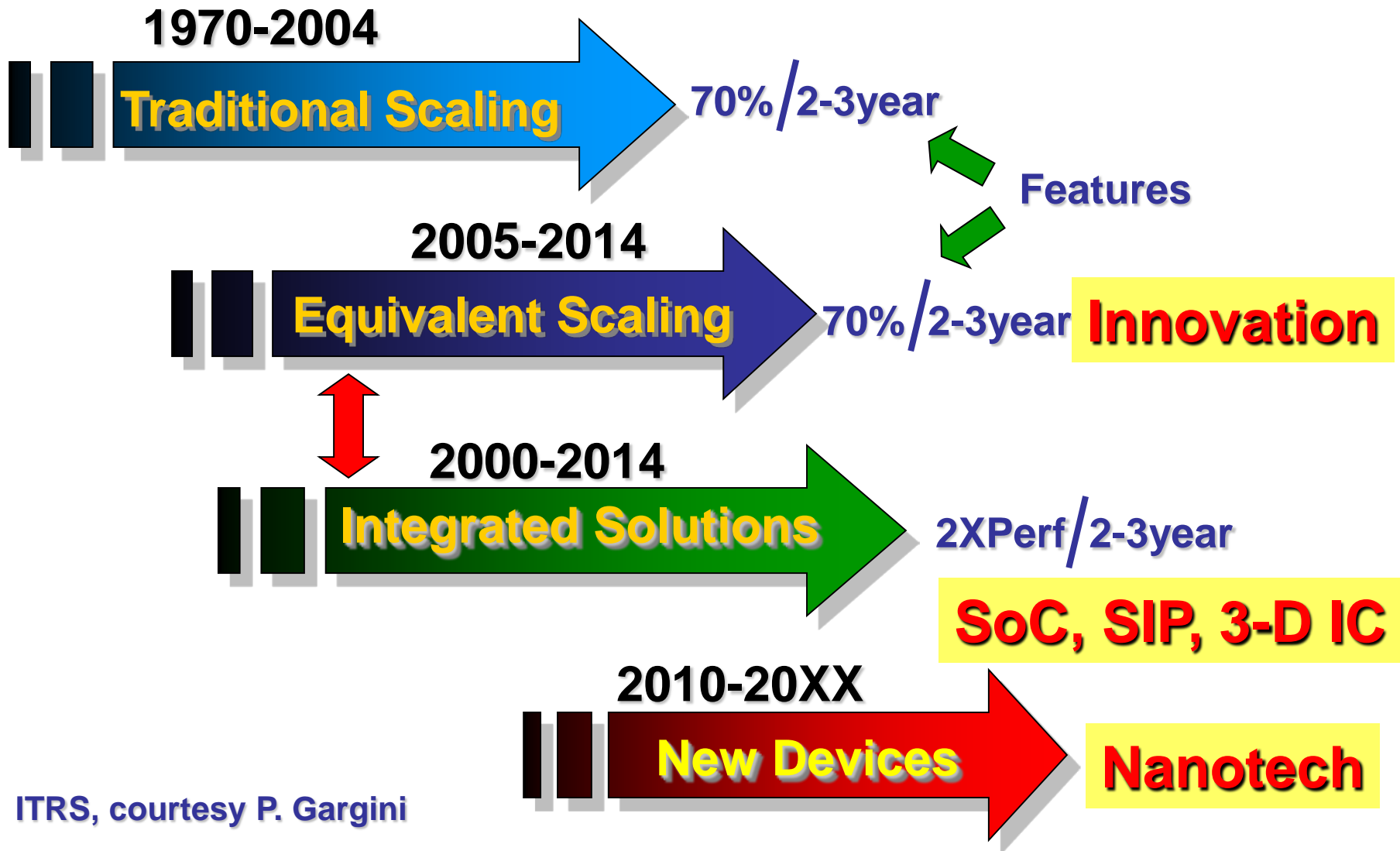
# Device Scaling



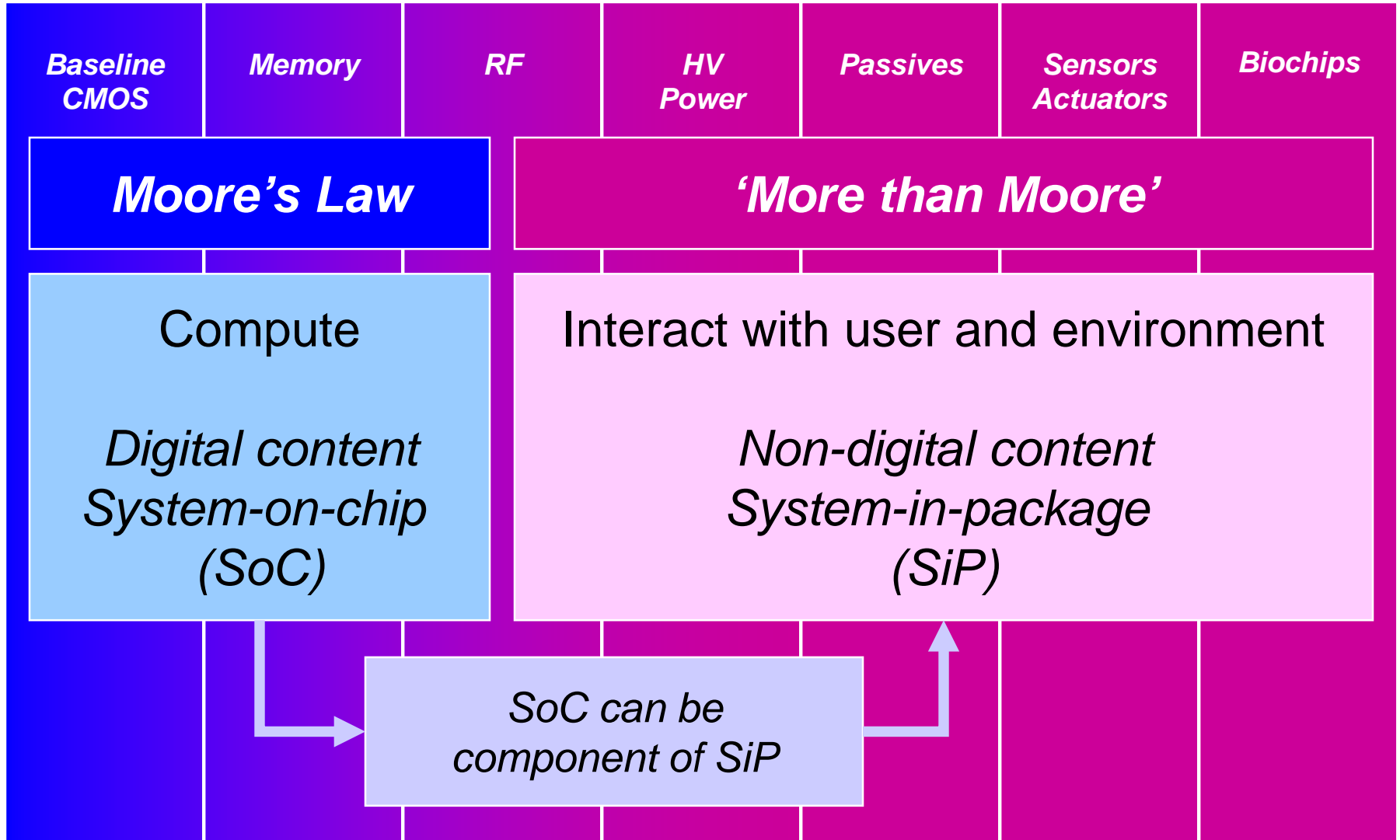
Source: Roger De Keersmaecker, IMEC

Lo/SVP

# CMOS Future Directions



# Moore's Law & More

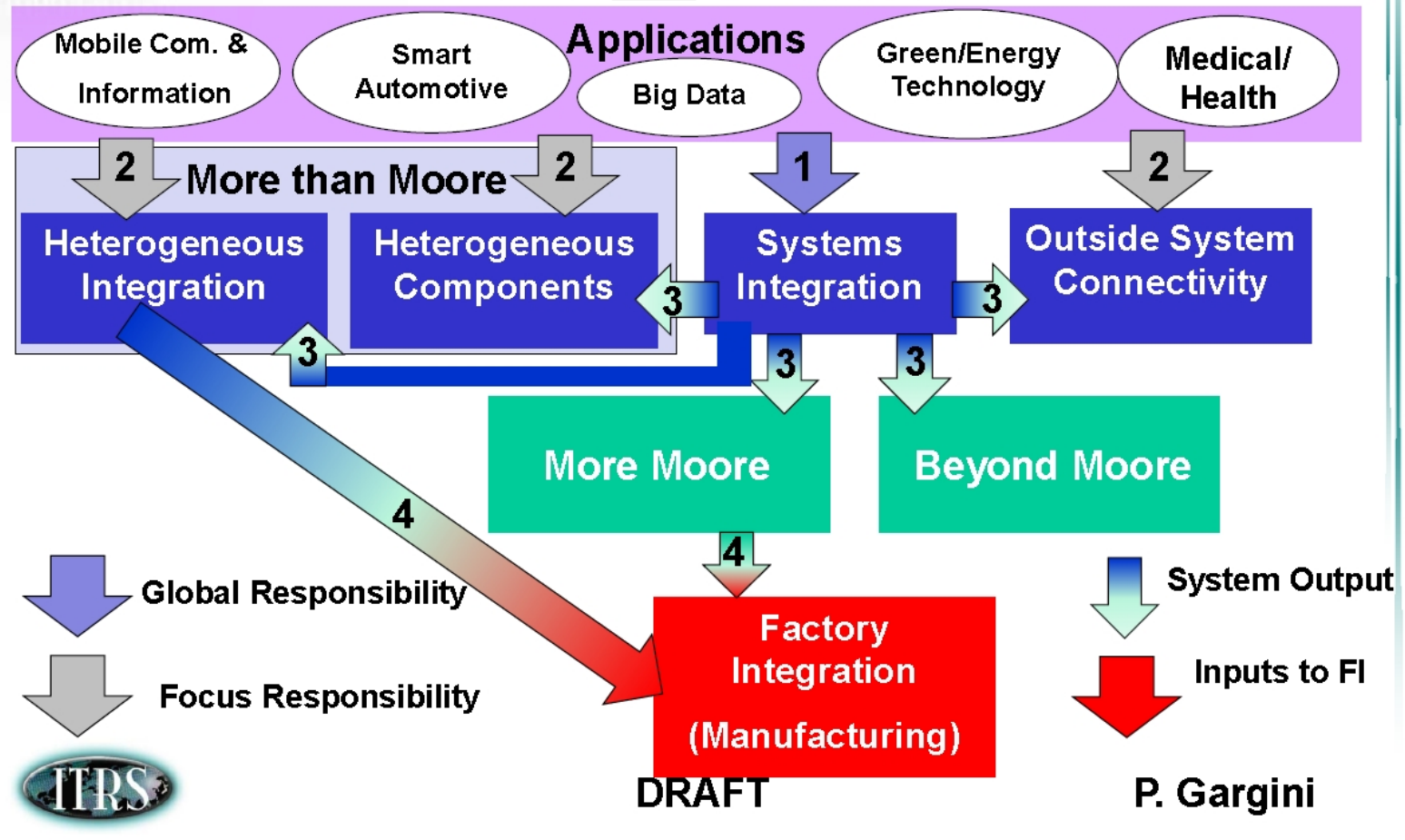


Source: ITRS 2006

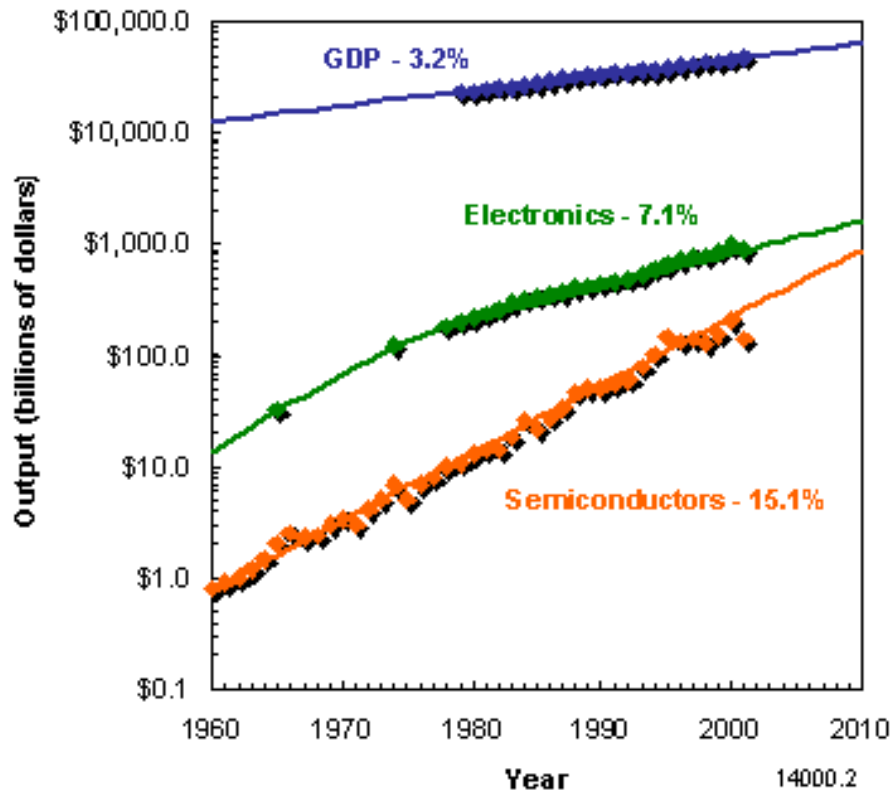
## Assessment of Driving Forces

# Internet of Things

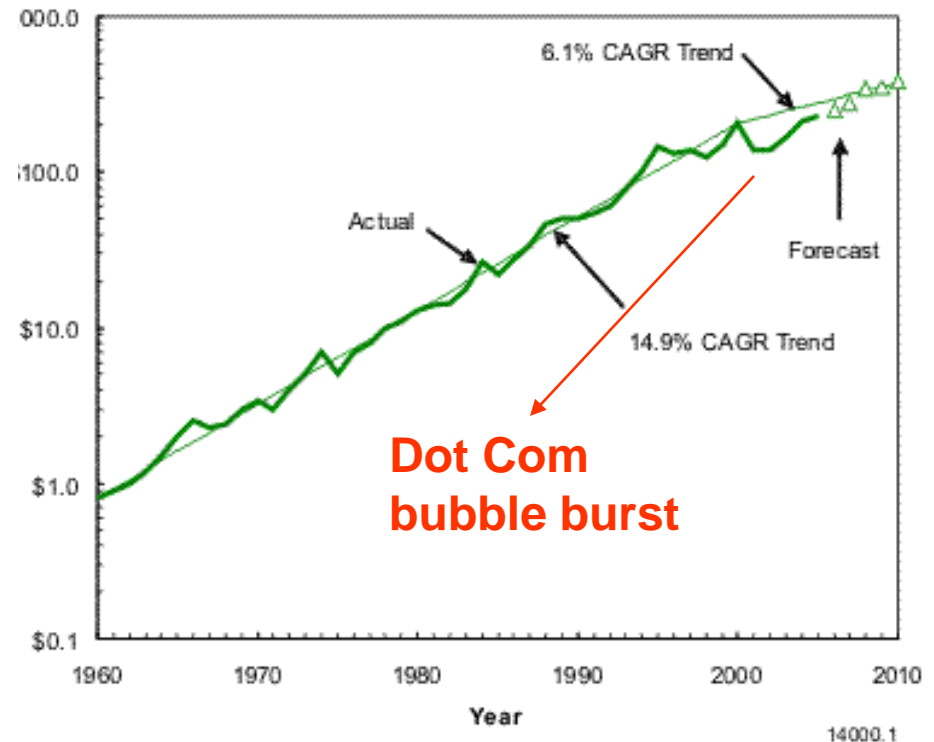
*Data Input, Access & Processing Environment*



# Semiconductor Growth 1



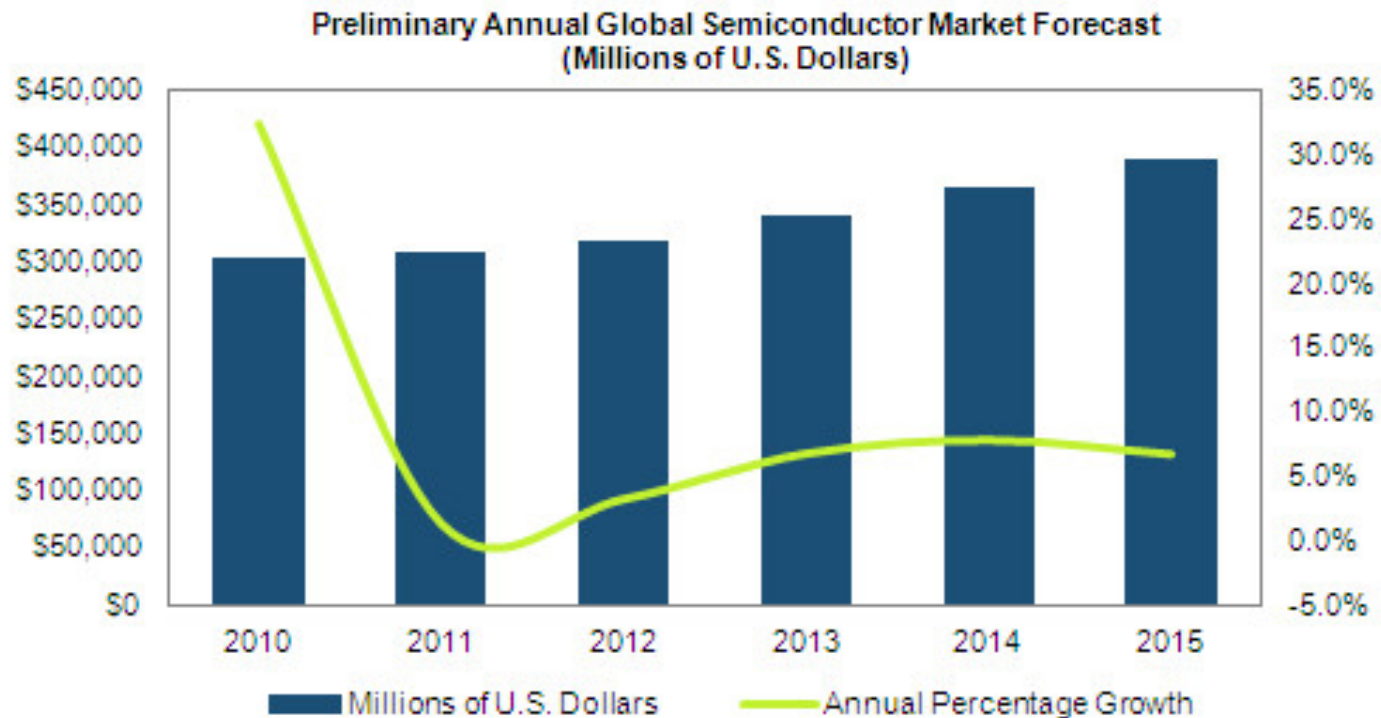
2002



Sept 2006

(IC Knowledge)

# Semiconductor Growth 2

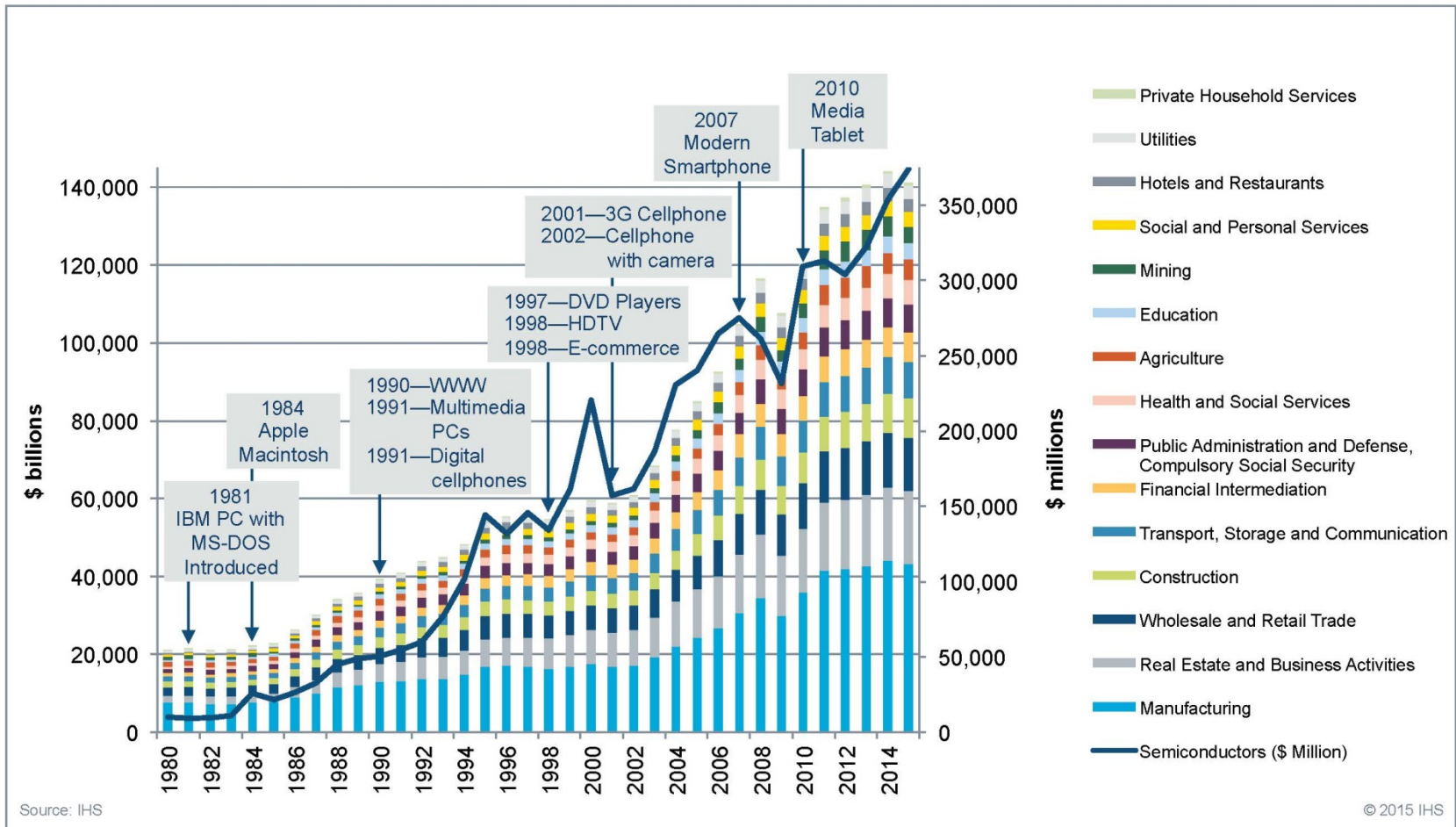


Source: IHS iSuppli November 2011



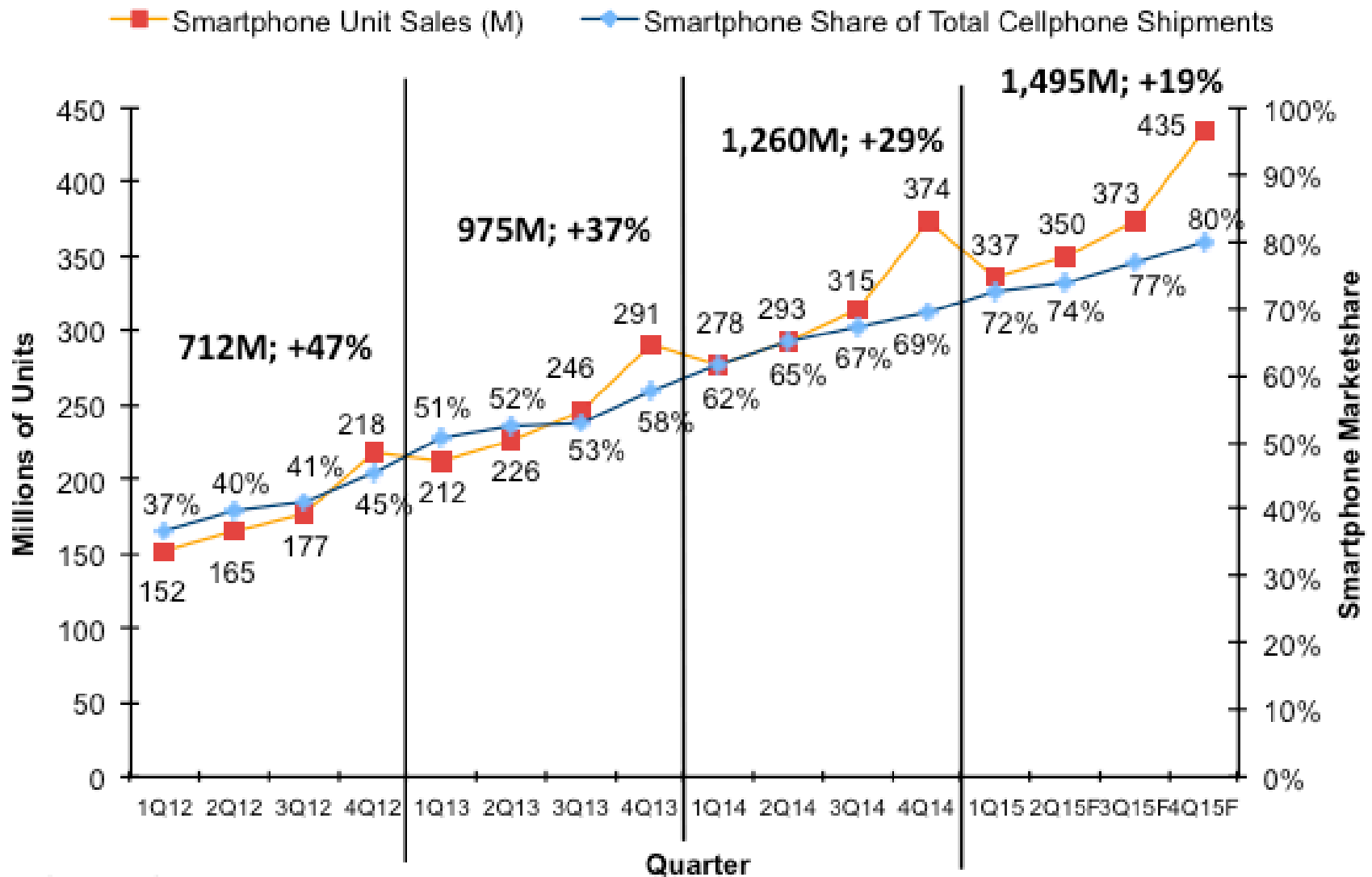
## Semiconductor growth tracks global output

Worldwide gross output by industry (\$ in billions, left axis) and semiconductor sales (\$ in millions, right axis)





## Smartphone Marketshare Trends



Source: IC Insights

# Semiconductor Revenue

World GDP in 2008 = US\$55T

Google Revenue in 2008 = US\$21.8B

Singapore GDP in 2008 = US\$181.9B

Microsoft Revenue in 2008 = US\$58.4B

IBM Revenue in 2008 – US\$103.6B

Qualcomm Revenue in 2009 – US\$10.4B

Intel Revenue in 2008 – US\$37.6B

TSMC Revenue in 2007 – US\$9.8B

**Integrated Device Manufacturers (IDM)** - Chip makers such as Intel that design, manufacture and sell their chips;

**Fab-less manufacturers (Design houses)** - Such as nVidia and Xilinx that design and sell chips but outsource manufacturing to foundry companies;

**Foundry Company** – Such as TSMC, UMC, and GF that manufacture chips designed and sold by their customers (Fab = Fabrication Plant);

**Fab-lites** – TI, Freescale;

**Outsourced Semiconductor Assembly and Test (OSAT)**

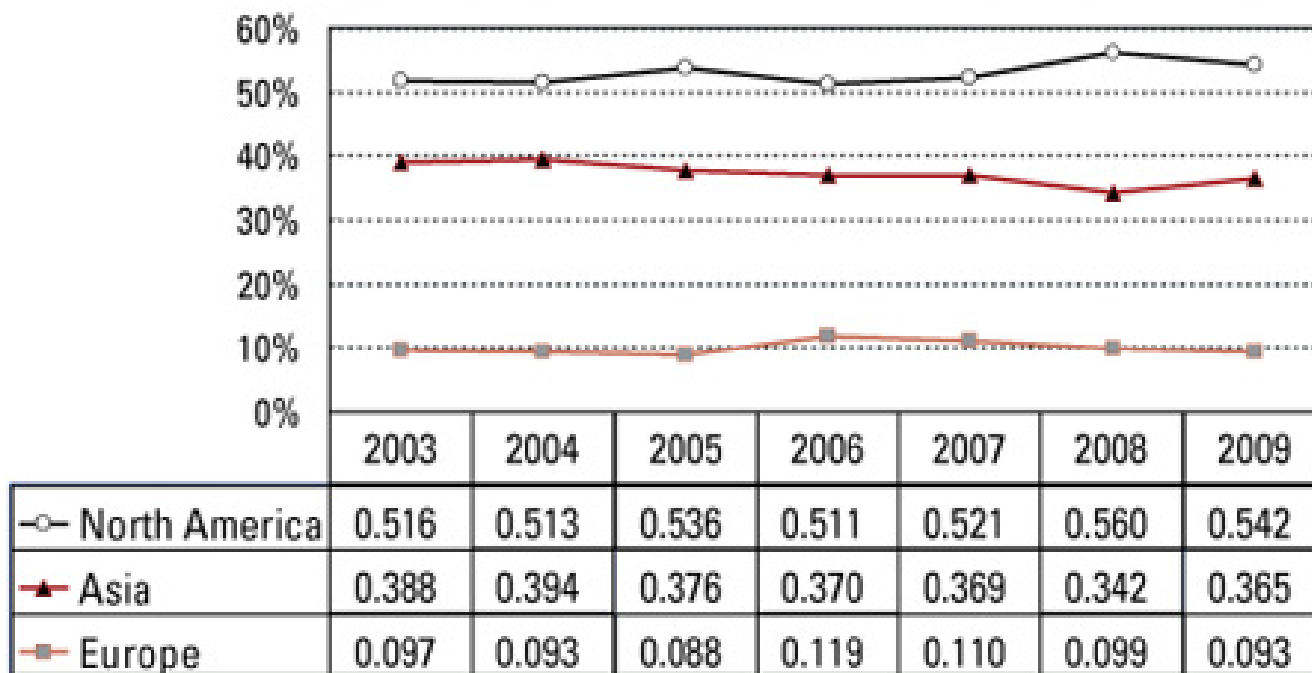
**Electronic Design Automation (EDA)**

**Equipments and Materials**

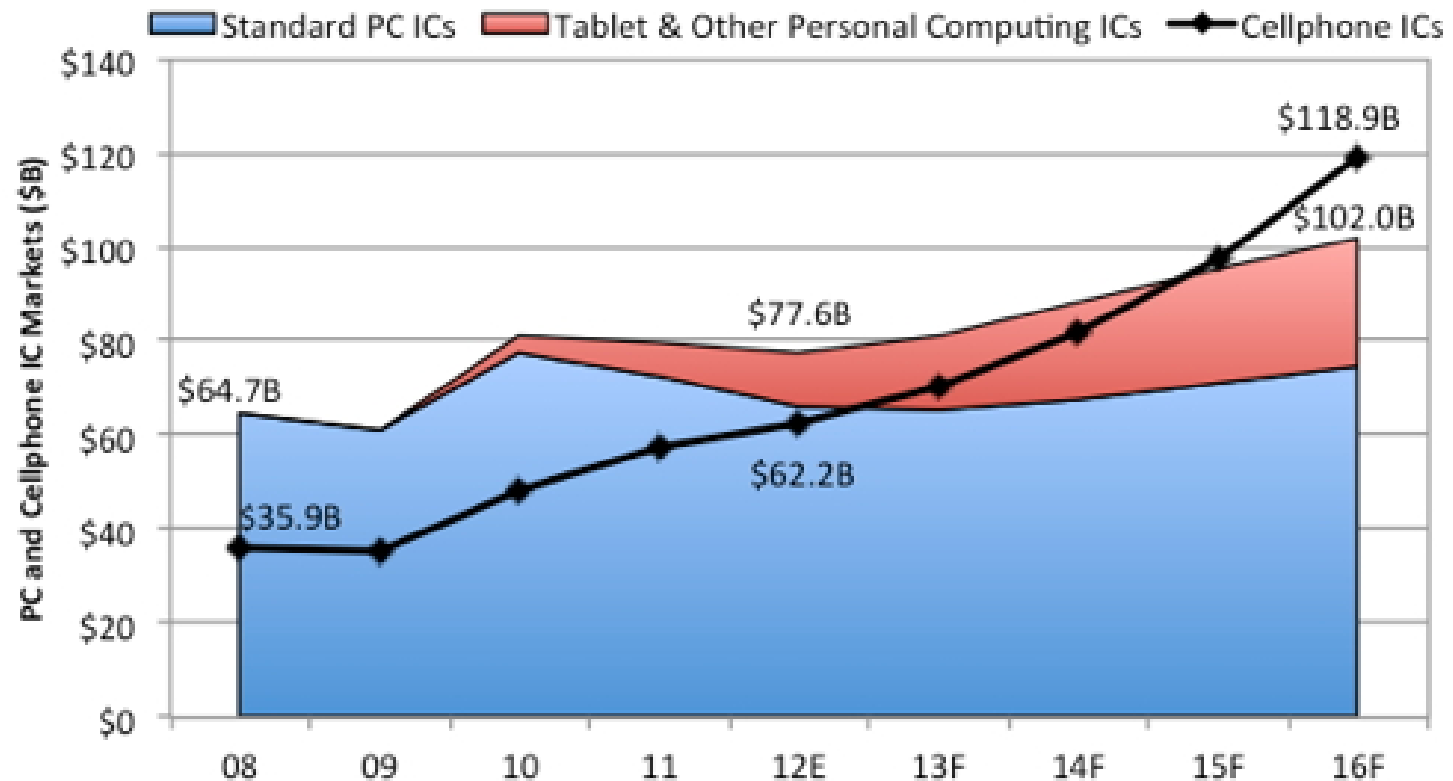
2012 Rank	2011 Rank	Company	Headquarters	2011 Tot IC	2011 Tot O-S-D	2011 Tot Semi	2012 Tot IC	2012 Tot O-S-D	2012 Tot Semi	2012/2011 % Change
1	1	Intel	U.S.	49,697	0	49,697	49,114	0	49,114	-1%
2	2	Samsung	South Korea	32,703	780	33,483	29,730	2,521	32,251	-4%
3	3	TSMC*	Taiwan	14,600	0	14,600	17,167	0	17,167	18%
4	7	Qualcomm**	U.S.	9,828	0	9,828	13,177	0	13,177	34%
5	4	TI	U.S.	12,182	718	12,900	11,442	705	12,147	-6%
6	5	Toshiba	Japan	10,024	2,721	12,745	9,055	2,162	11,217	-12%
7	6	Renesas	Japan	8,517	2,136	10,653	7,487	1,827	9,314	-13%
8	9	SK Hynix	South Korea	9,403	0	9,403	9,057	0	9,057	-4%
9	8	ST	Europe	7,117	2,514	9,631	6,227	2,137	8,364	-13%
10	10	Micron	U.S.	8,125	446	8,571	7,567	435	8,002	-7%
11	11	Broadcom**	U.S.	7,160	0	7,160	7,793	0	7,793	9%
12	13	Sony	Japan	4,706	1,387	6,093	4,449	1,260	5,709	-6%
13	12	AMD**	U.S.	6,568	0	6,568	5,422	0	5,422	-17%
14	14	Infineon	Europe	3,560	2,039	5,599	3,143	1,850	4,993	-11%
15	21	GlobalFoundries*	U.S.	3,480	0	3,480	4,560	0	4,560	31%
16	18	Nvidia**	U.S.	3,939	0	3,939	4,229	0	4,229	7%
17	15	Fujitsu	Japan	4,035	395	4,430	3,805	357	4,162	-6%
18	17	NXP	Europe	2,855	1,292	4,147	2,931	1,226	4,157	0%
19	16	Freescall	U.S.	3,750	641	4,391	3,164	571	3,735	-15%
20	20	UMC*	Taiwan	3,760	0	3,760	3,730	0	3,730	-1%
21	26	MediaTek**	Taiwan	2,969	0	2,969	3,366	0	3,366	13%
22	27	Sharp	Japan	1,658	1,250	2,908	1,799	1,505	3,304	14%
23	22	Marvell**	U.S.	3,445	0	3,445	3,157	0	3,157	-8%
24	19	Elpida	Japan	3,891	0	3,891	3,075	0	3,075	-21%
25	24	Rohm	Japan	1,952	1,351	3,303	1,792	1,238	3,030	-8%
Top 25 Total				219,924	17,670	237,594	216,438	17,794	234,232	-1%

\*Foundry

\*\*Fabless



## IC Market for PCs vs. Cellphones



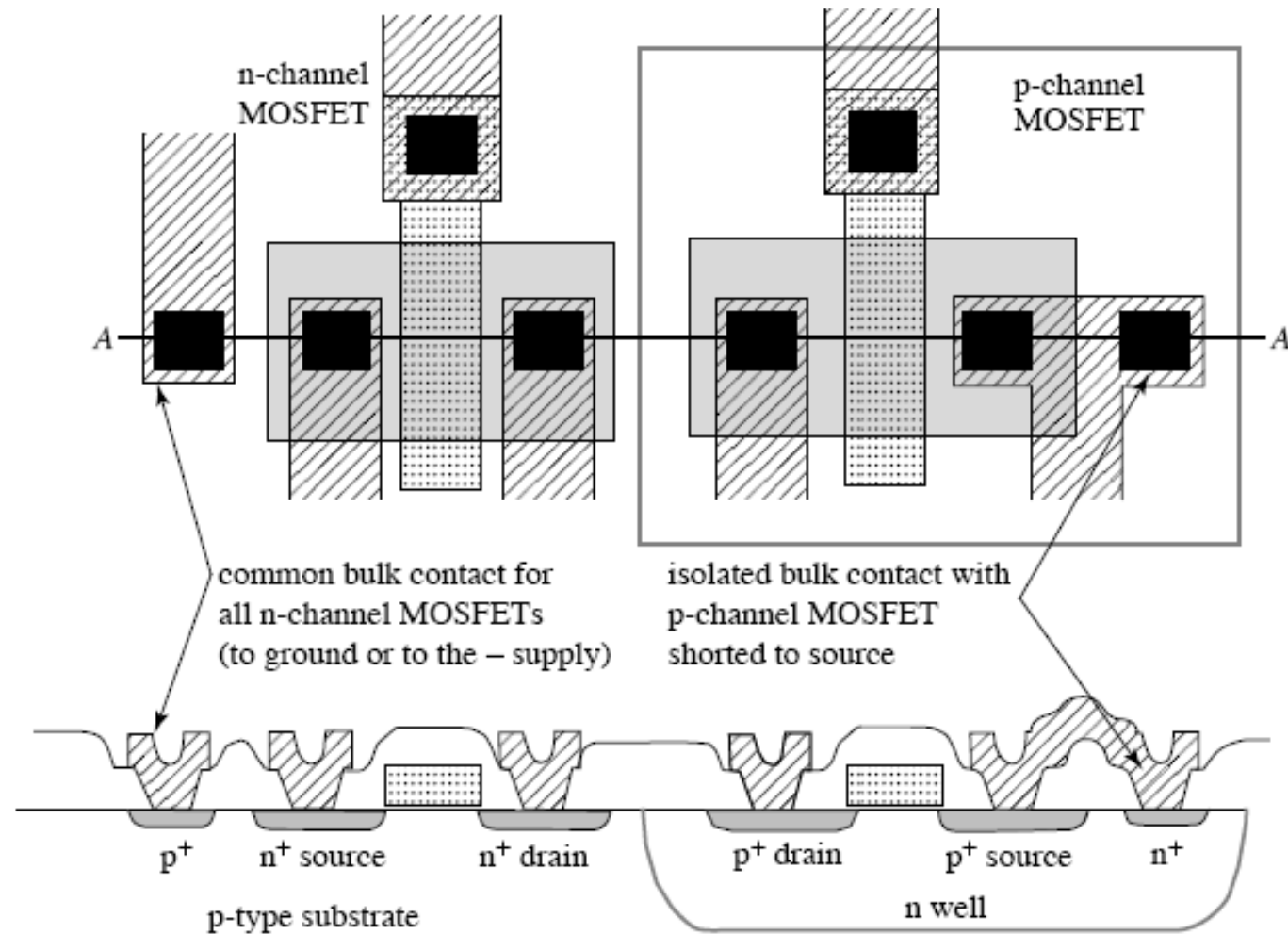
Source: IC Insights

# IC Technology Market Share (2000)

- CMOS 77%
- Analog Bipolar 6%
- Digital Bipolar 2%
- BiCMOS 6%
- GaAs 6%
- Other 3%

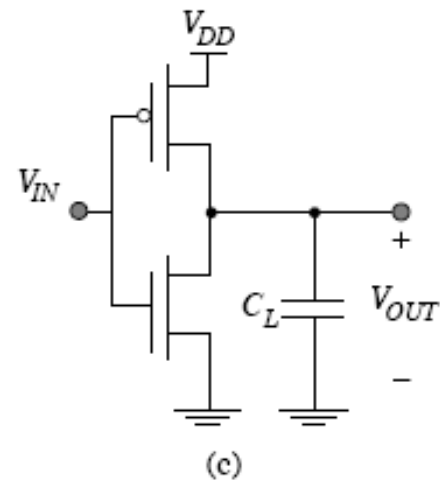
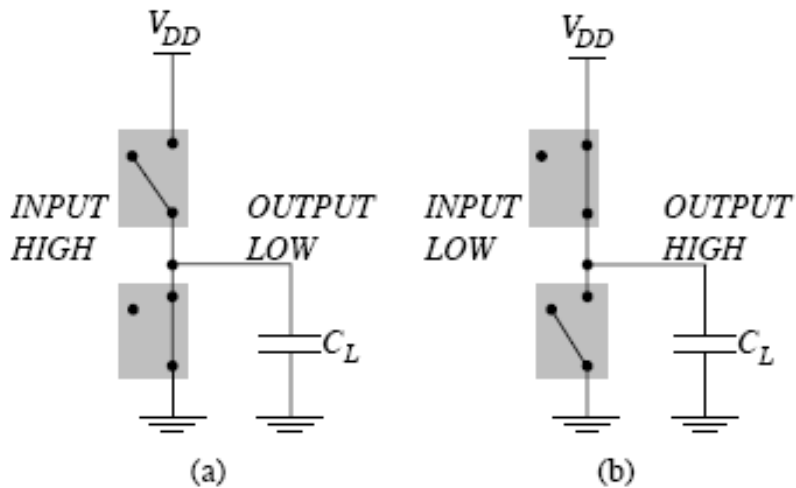
**Silicon-based**

# Complementary Metal Oxide Semiconductor (CMOS)



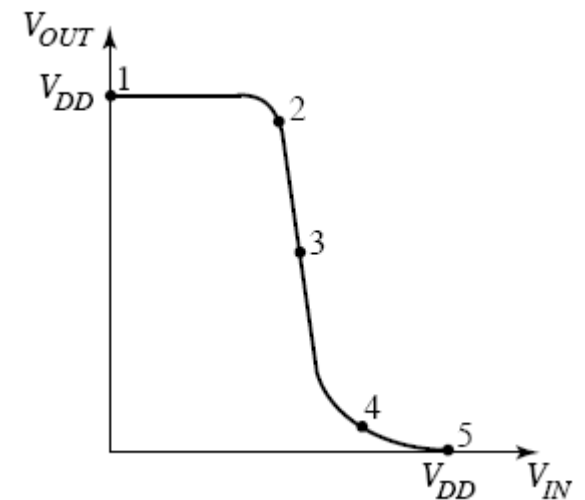
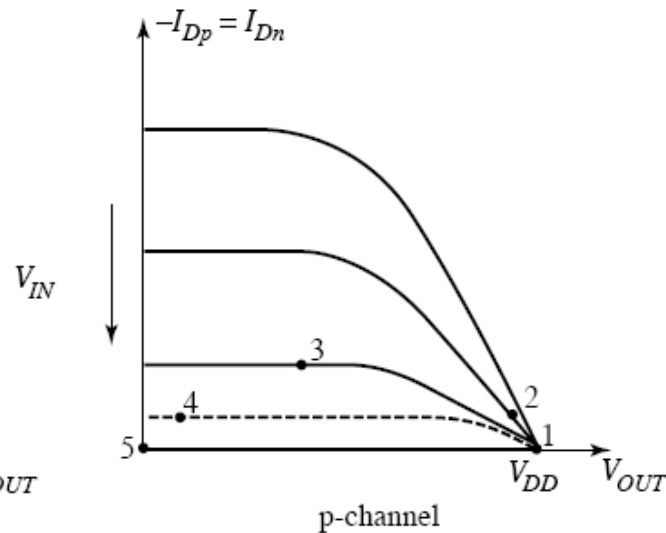
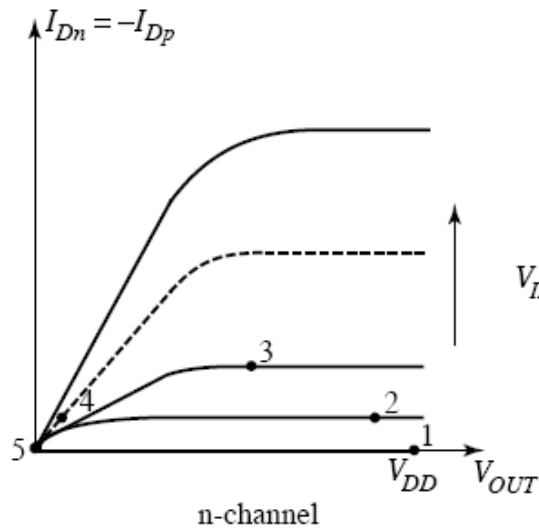
Howe et al, Prentice Hall

# CMOS Operation



A	B
0	1
1	0

CMOS Inverter



Howe et al, Prentice Hall

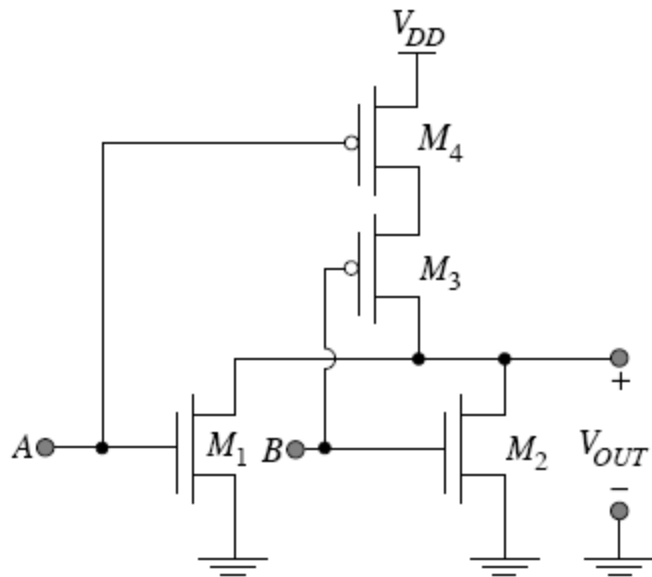


# CMOS Logic Gates

NOR



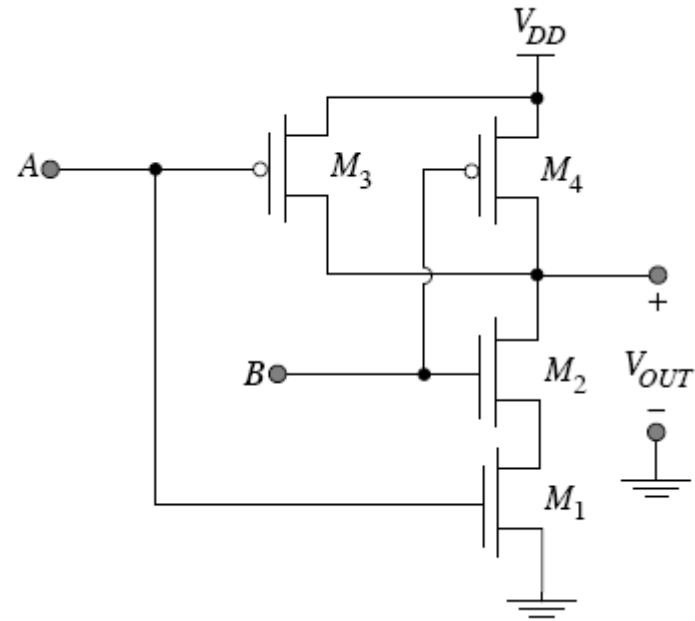
INPUT		OUTPUT
A	B	A NOR B
0	0	1
0	1	0
1	0	0
1	1	0



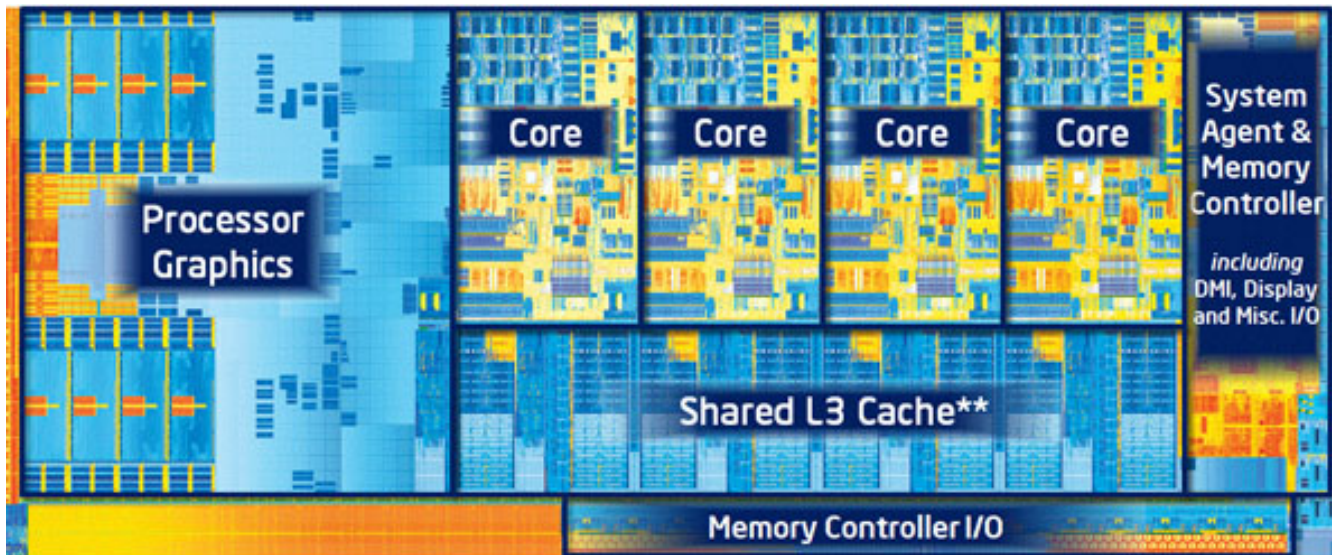
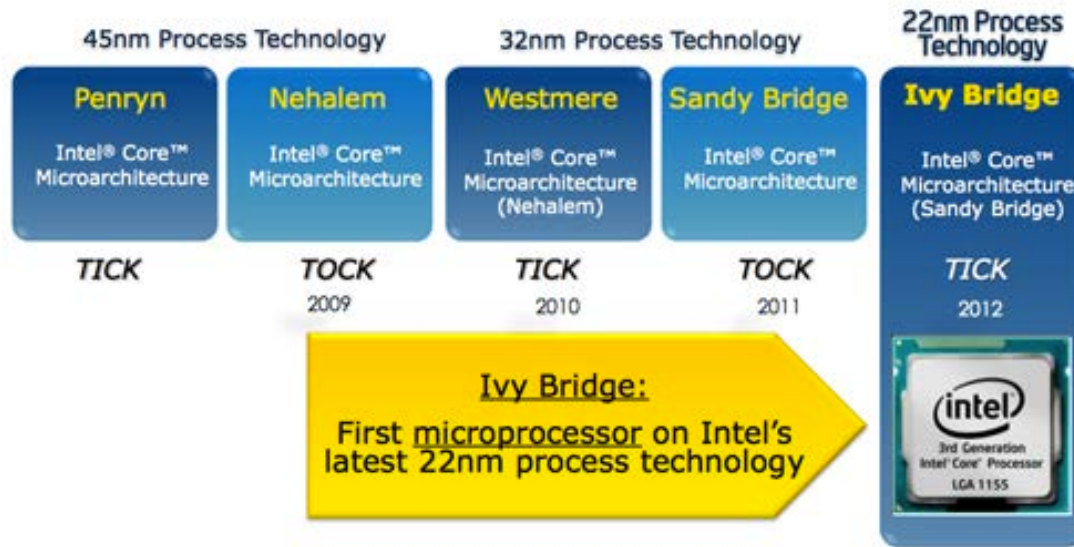
NAND



Input A	Input B	Output
0	0	1
0	1	1
1	0	1
1	1	0



Howe et al, Prentice Hall



# Process Technology Leadership

2003  
90 nm



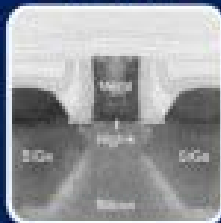
Invented  
SiGe  
Strained Silicon

2005  
65 nm



2nd Gen.  
SiGe  
Strained Silicon

2007  
45 nm



Invented  
Gate-Last  
High-k  
Metal Gate

2009  
32 nm



2nd Gen.  
Gate-Last  
High-k  
Metal Gate

2011  
22 nm



First to  
Implement  
Tri-Gate

STRAINED SILICON

HIGH-k METAL GATE

TRI-GATE

## 22nm

A Revolutionary  
Leap in  
Process Technology

### 37%

Performance Gain at Low  
Voltage\*

### >50%

Active Power Reduction at  
Constant Performance\*

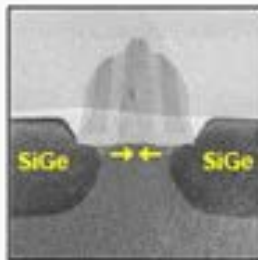
*The foundation for all computing*



Source: Intel  
\*Compared to Intel 32nm Technology

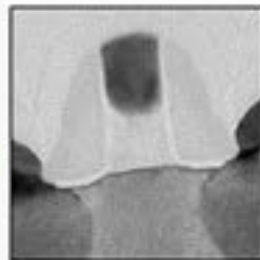
16

2003  
90 nm



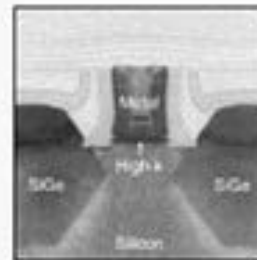
Invented  
SiGe  
Strained Silicon

2005  
65 nm



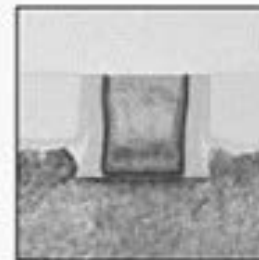
2<sup>nd</sup> Gen.  
SiGe  
Strained Silicon

2007  
45 nm



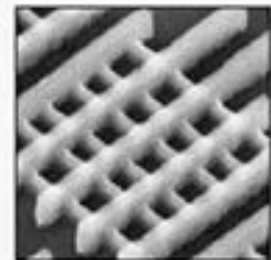
Invented  
Gate-Last  
High-k  
Metal Gate

2009  
32 nm



2<sup>nd</sup> Gen.  
Gate-Last  
High-k  
Metal Gate

2011  
22 nm



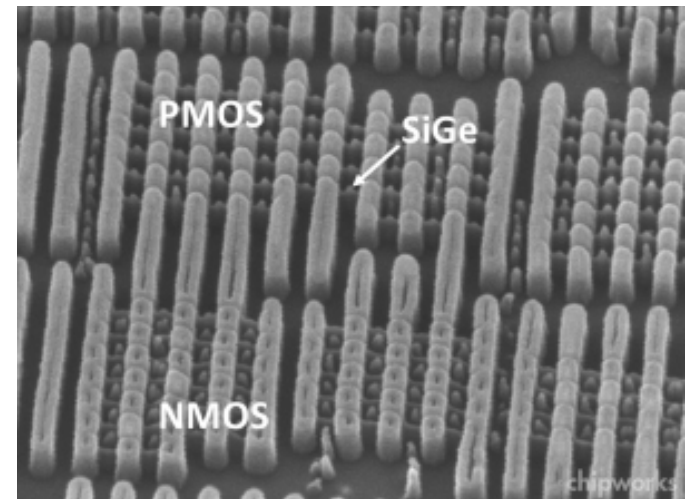
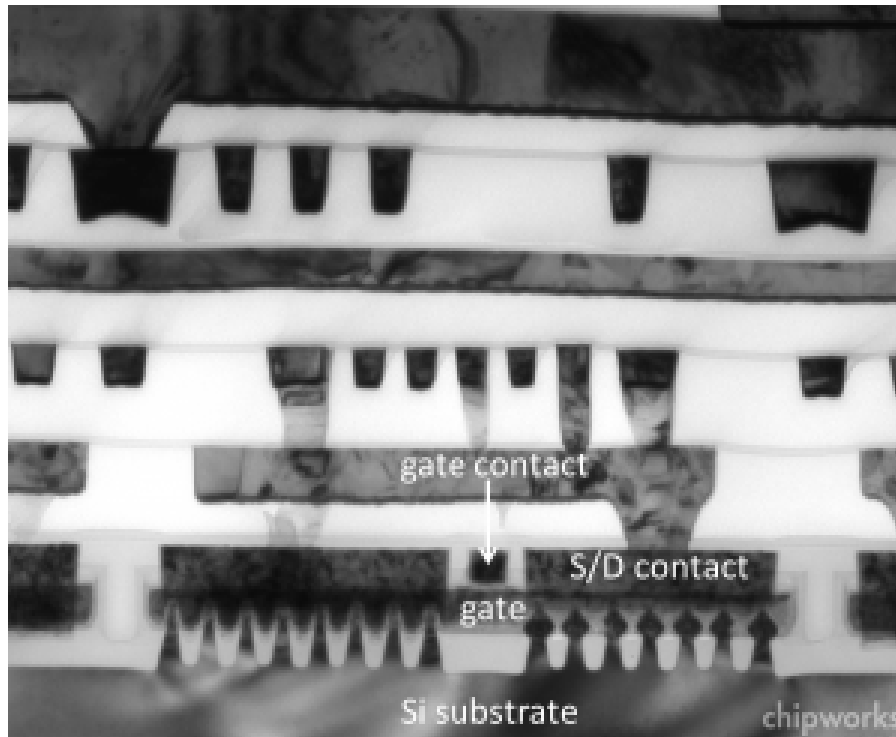
First to  
Implement  
Tri-Gate

Strained Silicon

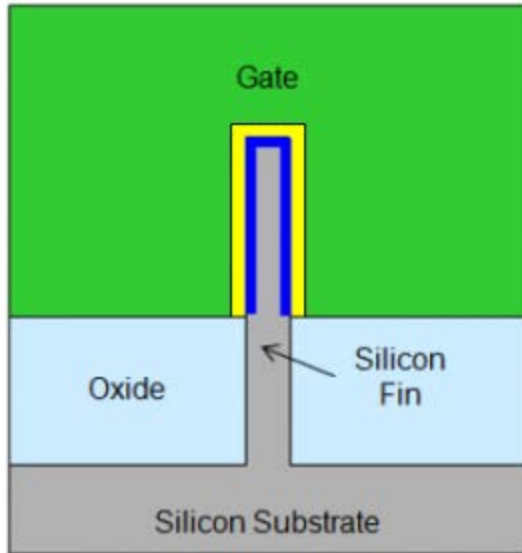
High-k Metal Gate

Tri-Gate

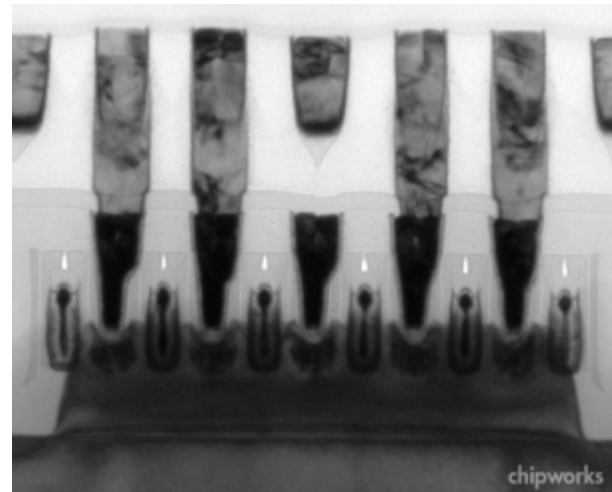
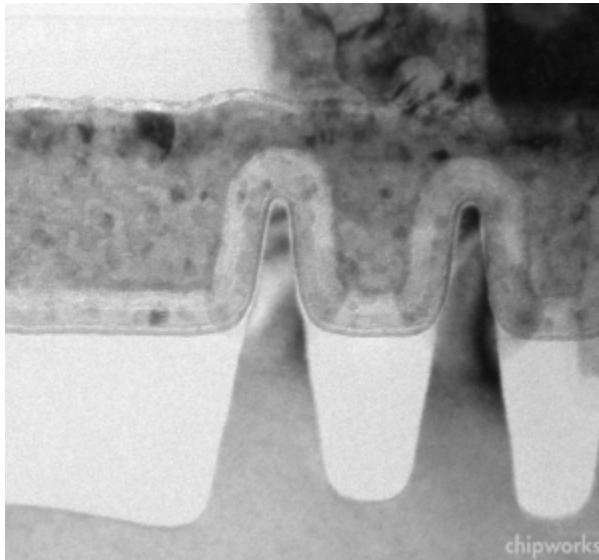
# Intel 22nm



TEM Image of Lower Metals and NMOS and PMOS (right) Transistors

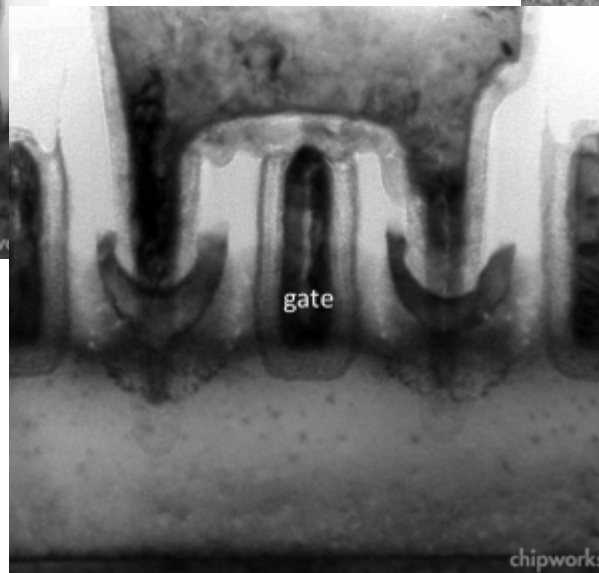
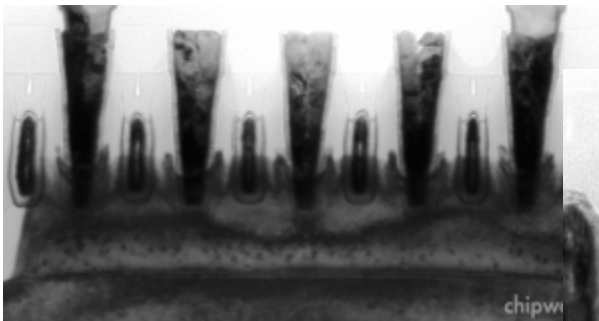
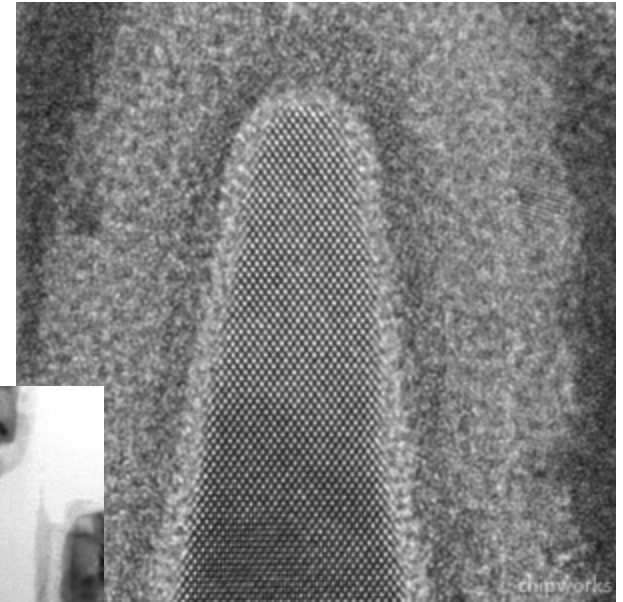
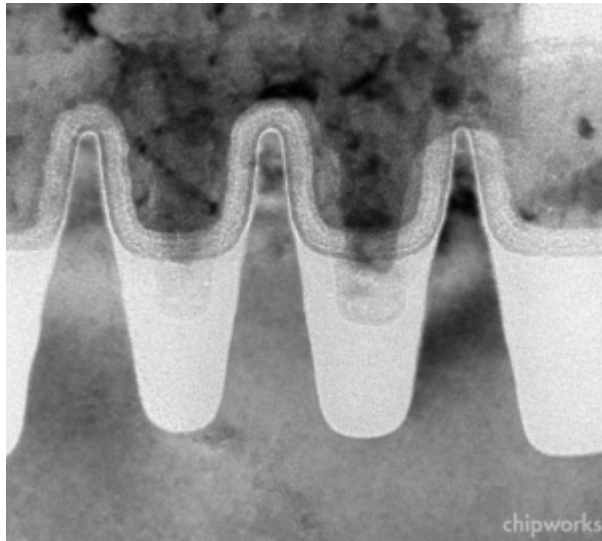


TEM Image of PMOS Gate and Fin Structure

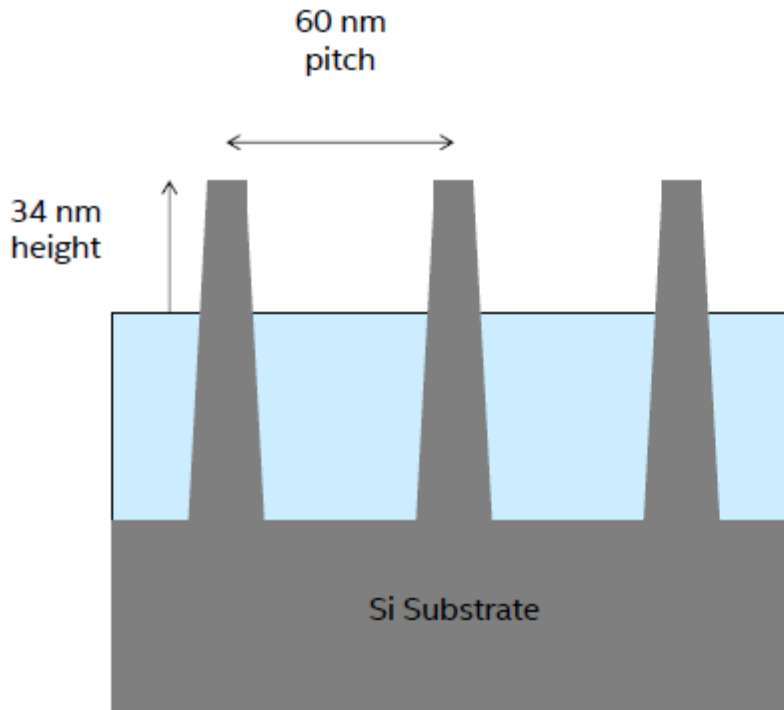




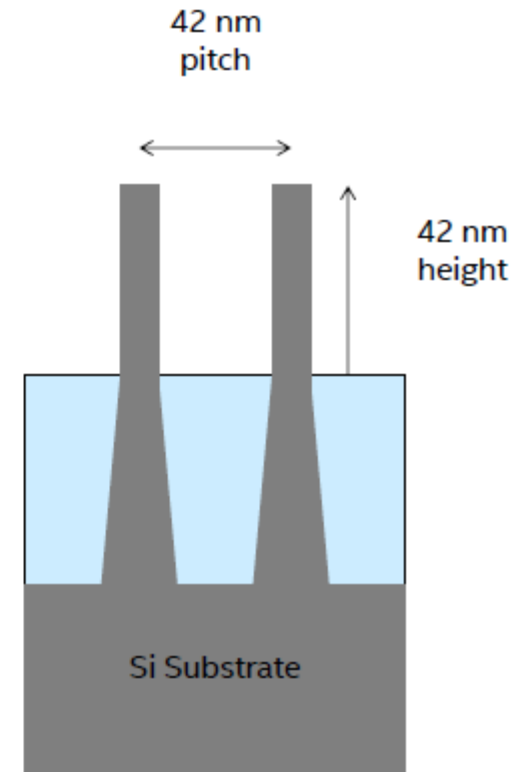
## TEM Image of NMOS Gate and Fin Structure



# Transistor Fin Improvement



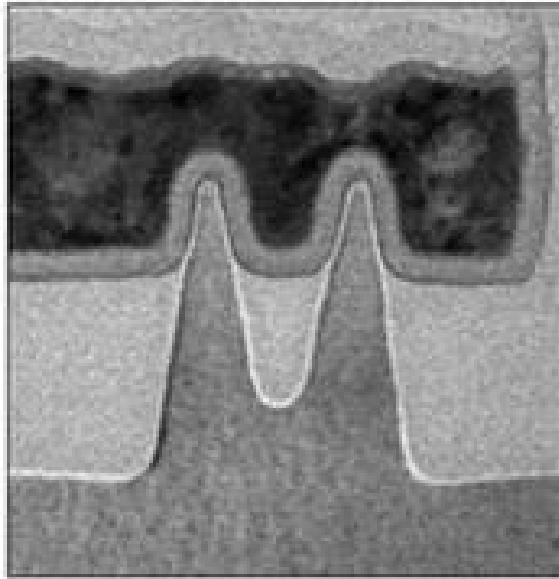
22 nm Process



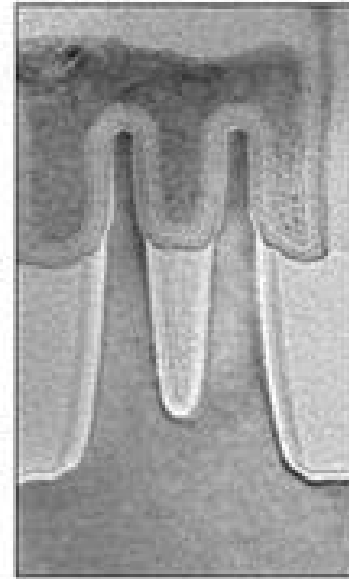
14 nm Process



# Transistor Fin Improvement



22 nm 1<sup>st</sup> Generation  
Tri-gate Transistor



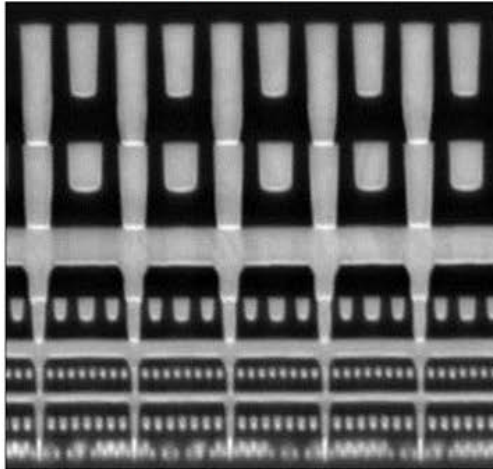
14 nm 2<sup>nd</sup> Generation  
Tri-gate Transistor

The size of transistor gates and "fins," especially to interconnection, were reduced by more than a third from the previous generation of technology.

(Image: Intel 2014)

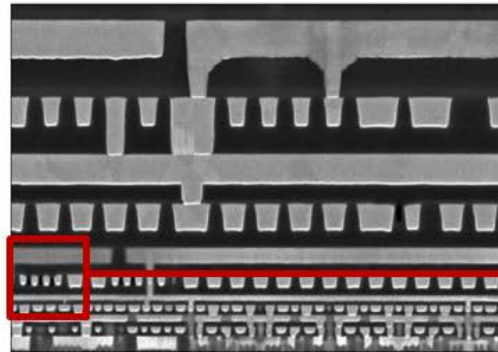
## Interconnects

22 nm Process

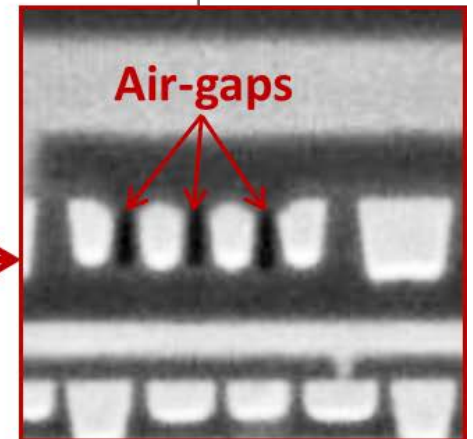


80 nm minimum pitch

14 nm Process



52 nm (0.65x) minimum pitch



*52 nm Interconnect Pitch Provides  
Better-than-normal Interconnect Scaling*



23

17% improvement  
in RC delay