NBAY 6120

Chip Making

&

Microprocessor Technology

NBAY 1620

March 2, 2017

Donald P. Greenberg

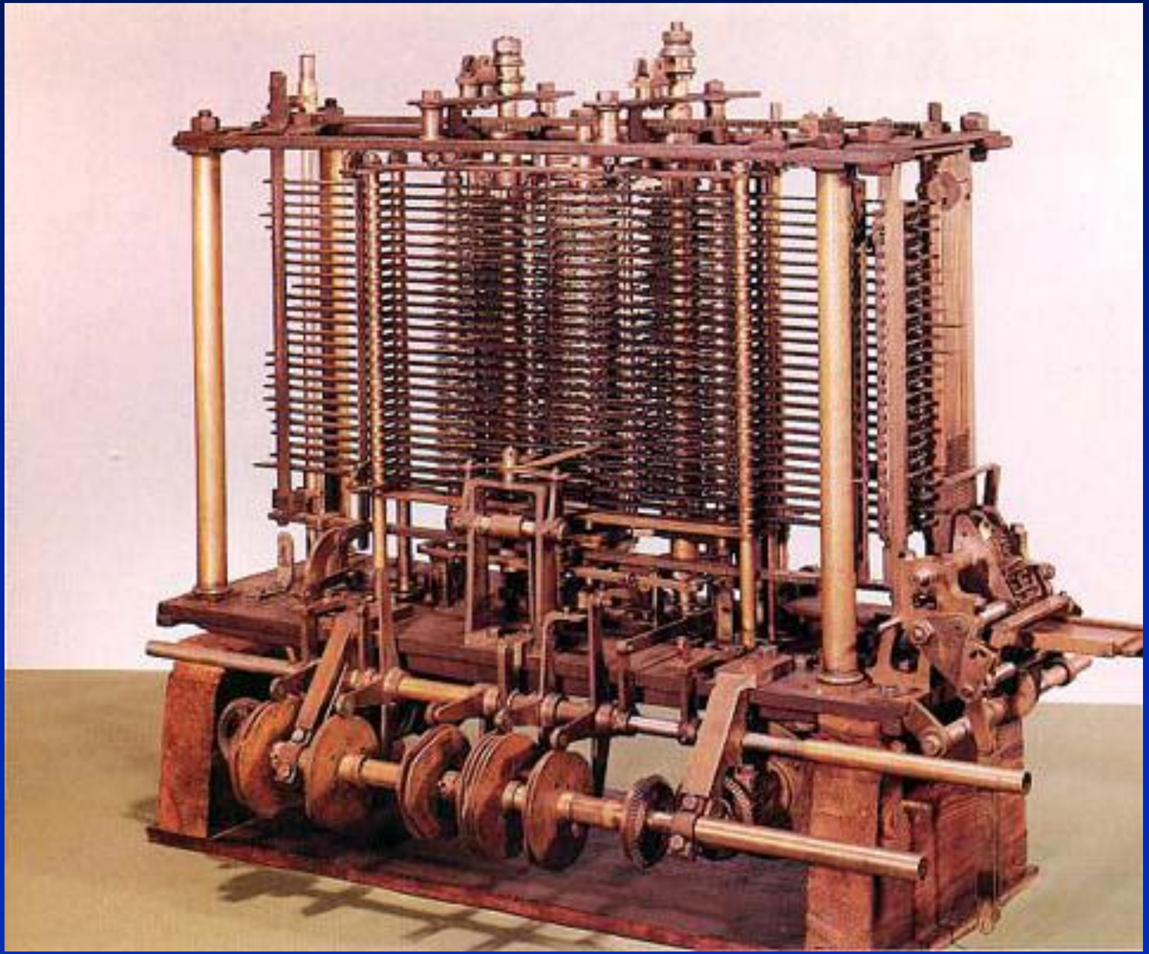
Lecture 2

- ***Required Reading:***
  - Craig R. Barrett. From Sand to Silicon: Manufacturing an Integrated Circuit, Scientific American, Special Issue, The Solid-State Century, January 1998, pp. 55-61. (Search: e-Journals/ Scientific American Archive Online/article (full text) http://www.library.cornell.edu/johnson/library/general/emba.html
  - Peter J. Denning and Ted G. Lewis. “Exponential Laws of Computing Growth.” Communications of the ACM. January 2017. ACM.org.

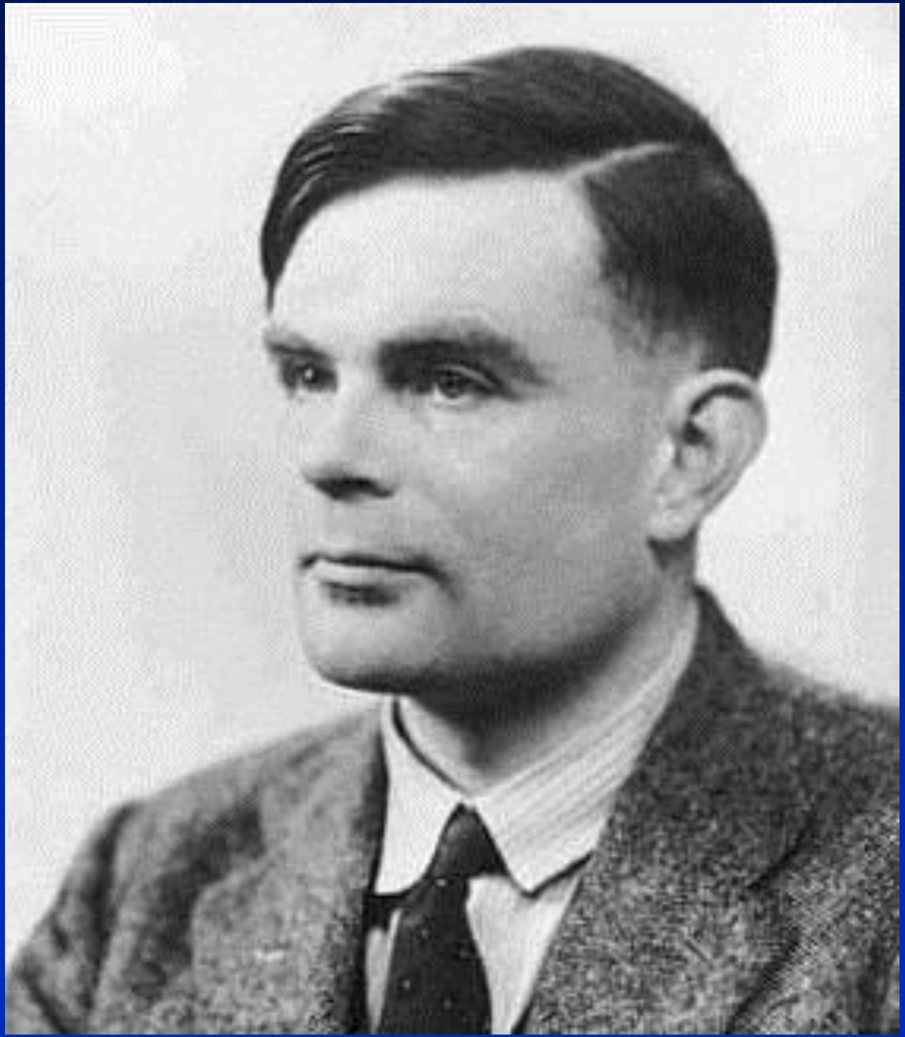
Optional:

Mack, Chris. "The Multiple Lives of Moore's Law." *IEEE Spectrum* Apr. 2015: 30-37. *Cornell University Library*. Web. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7065415

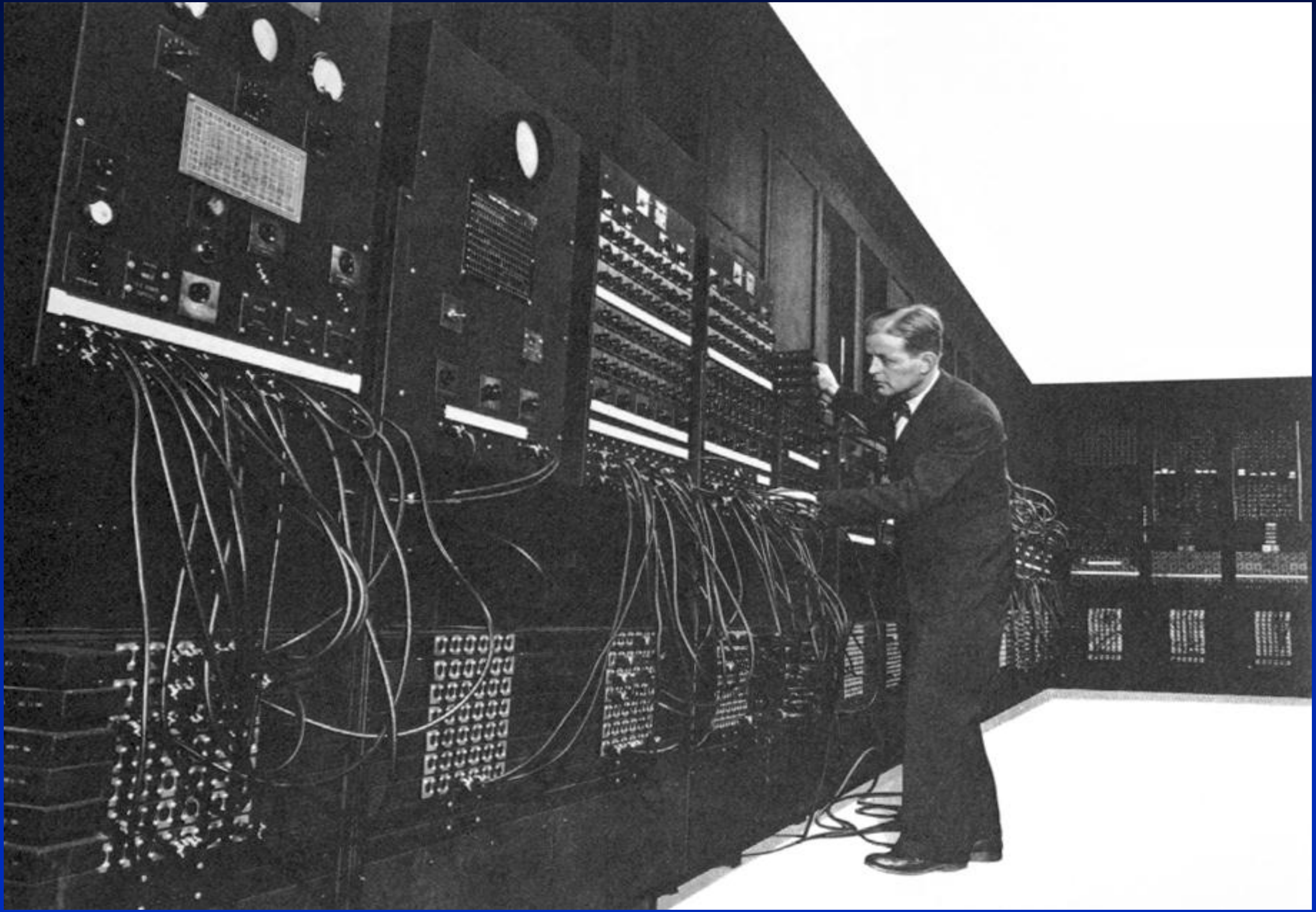
Turing Machine



Alan Turing

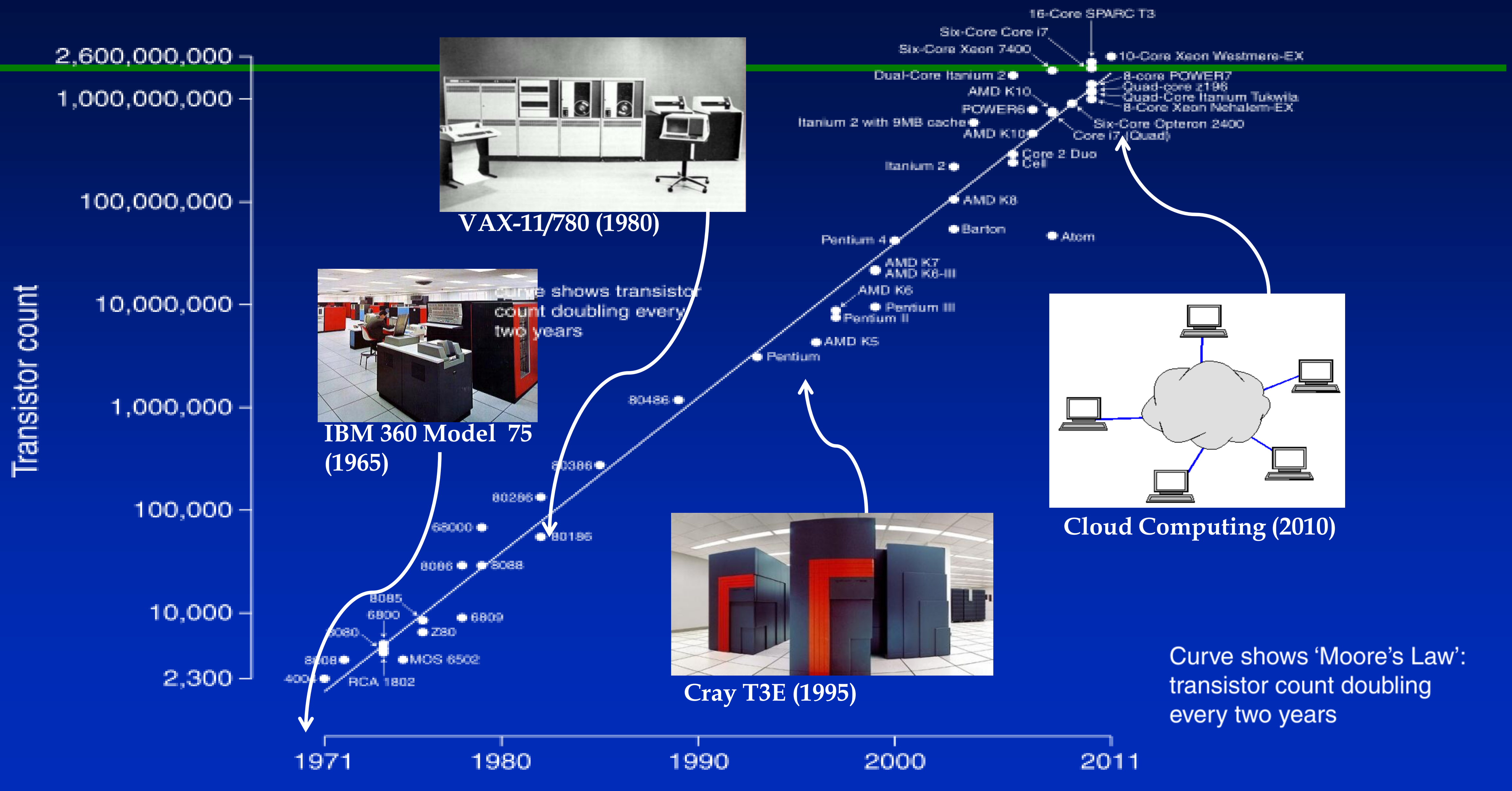


Eniac 1946



Cloud Computing - 2010

Microprocessor Transistor Counts 1971-2011 & Moore's Law



VAX-11/780 (1980)

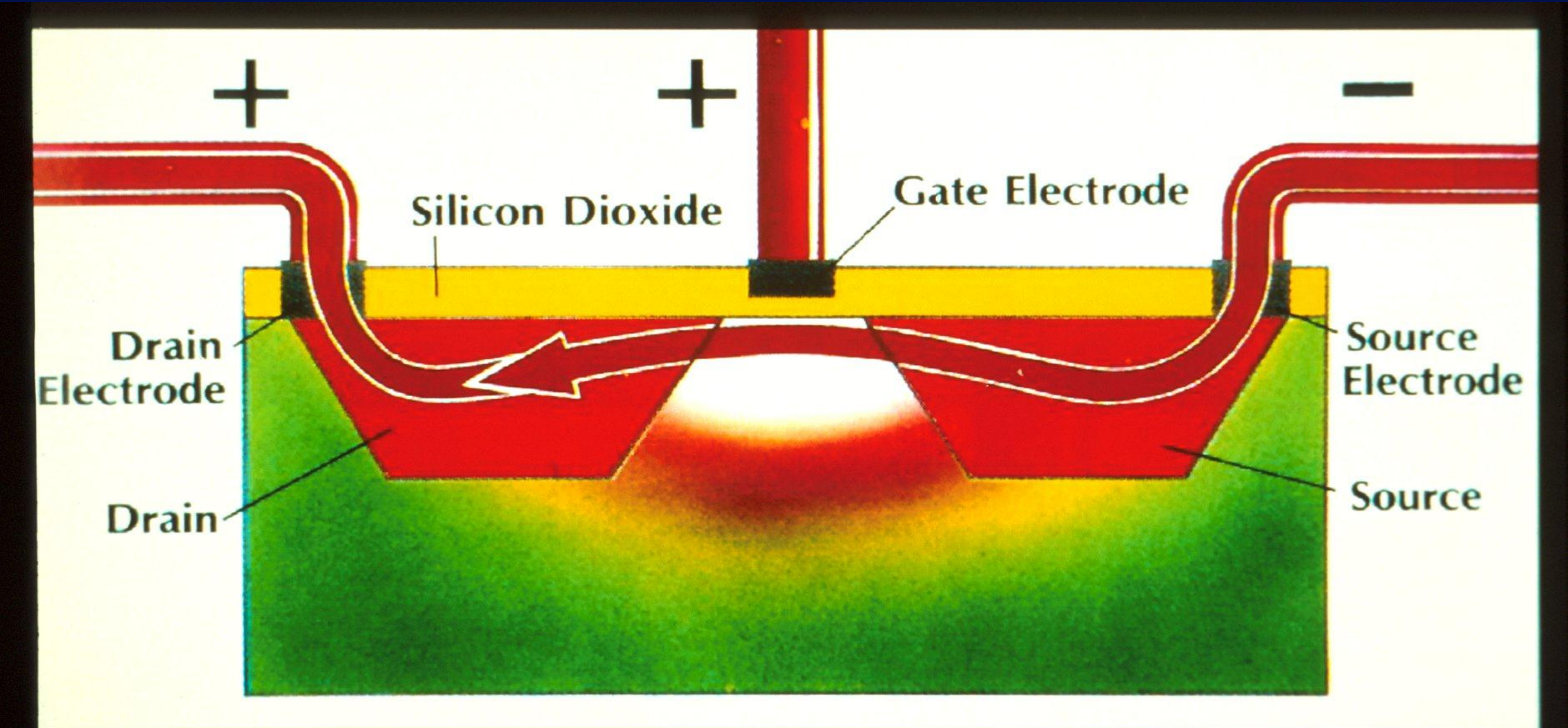
IBM 360 Model 75 (1965)

Cray T3E (1995)

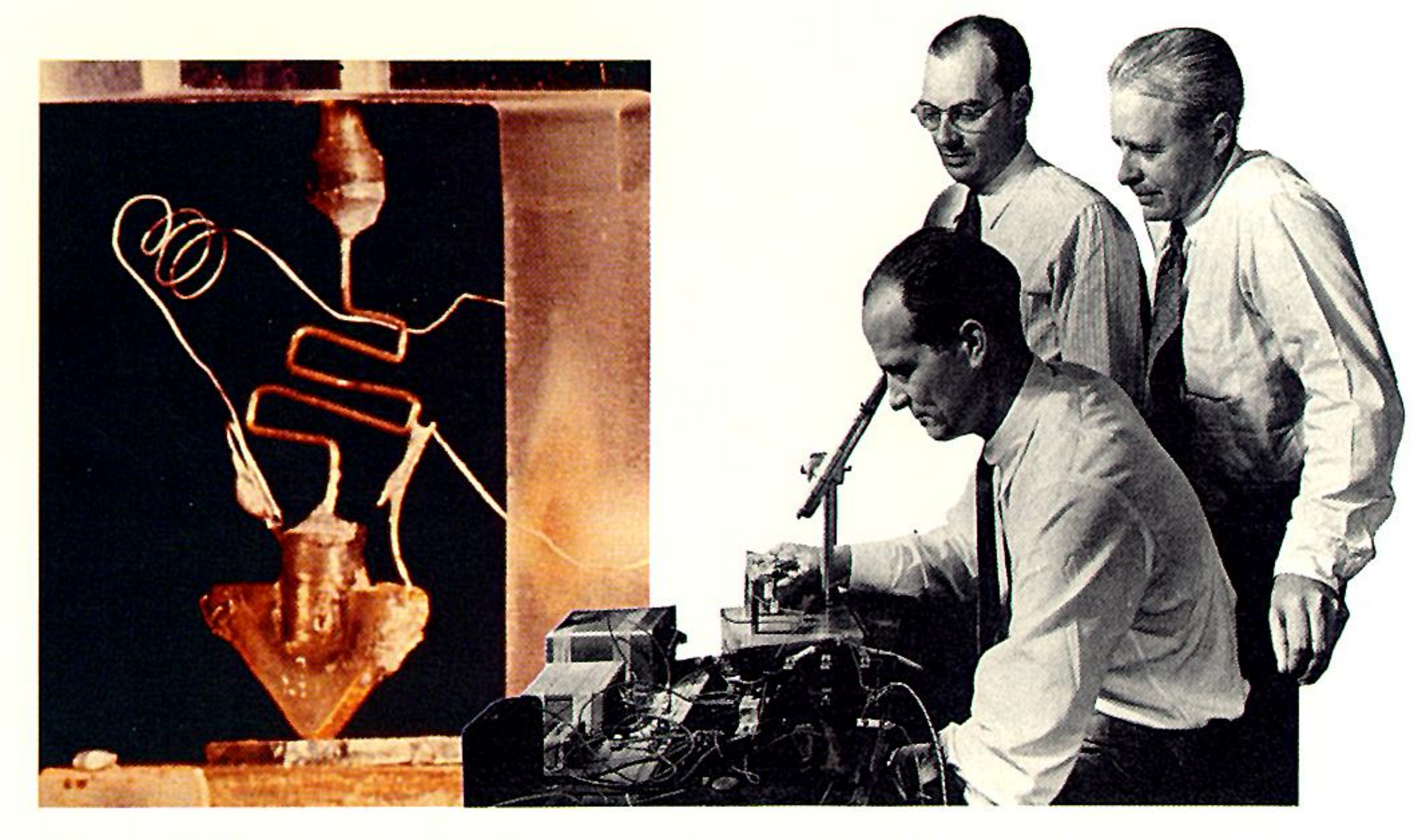
Cloud Computing (2010)

Curve shows 'Moore's Law': transistor count doubling every two years

Transistor


+
+
−
Silicon Dioxide
Gate Electrode
Drain Electrode
Source Electrode
Drain
Source

Shockley, Bardeen & Brattain 1947



Shockley & the Traitorous Eight

- William Shockley - Receives the Nobel Prize in Physics with Bardeen and Brattain (1956)
  leaves Bell Laboratory and forms Fairchild Semiconductor
- Julius Blank - founded Xicor
- Jean Hoerni - invented the planar process
  founded Amelco → Teledyne
- Jay Last - founded Amelco → Teledyne
- Sheldon Roberts - founded Amelco → Teledyne

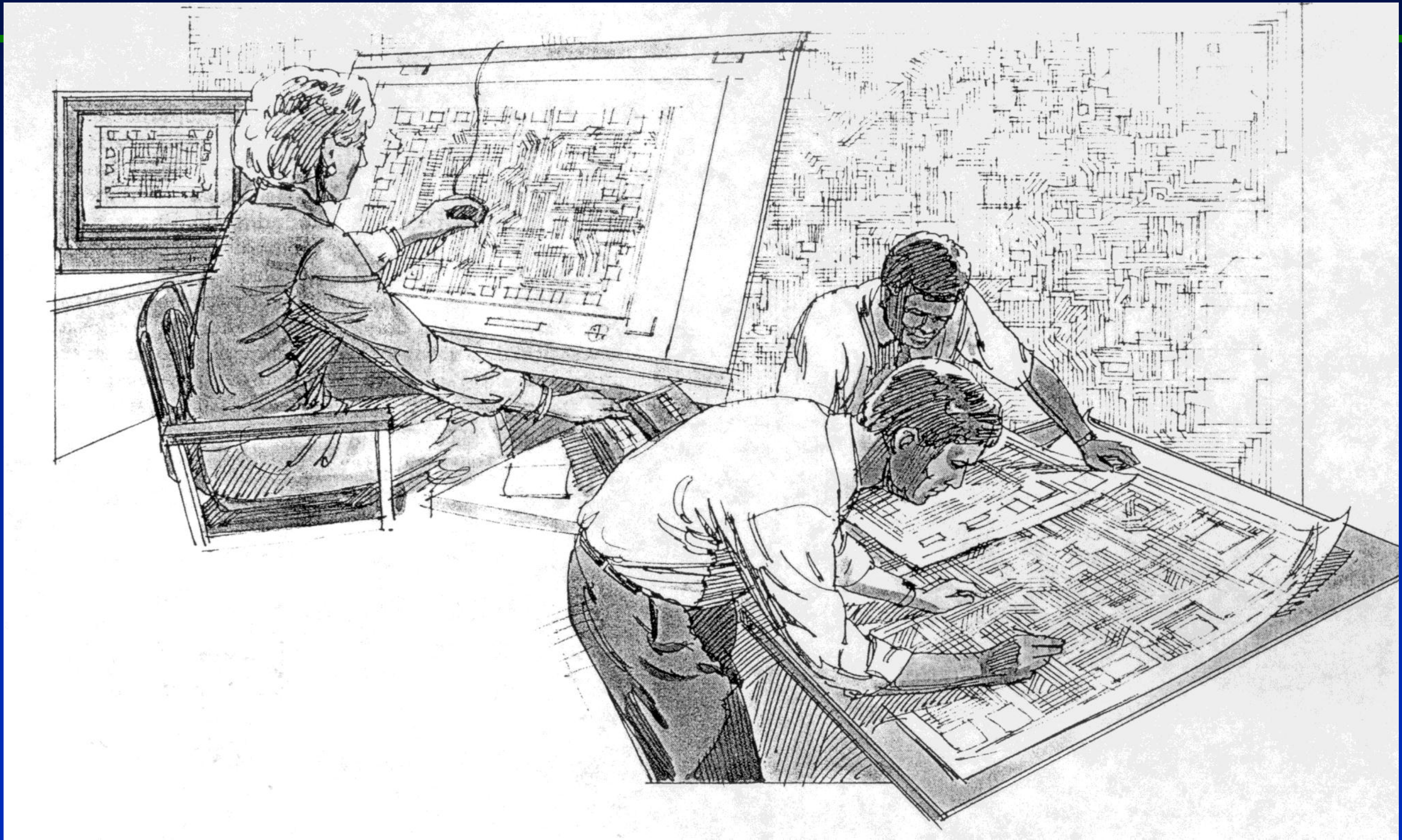
Shockley & the Traitorous Eight

• Gordon Moore -	founded Intel in 1968
• Robert Noyce -	founded Intel in 1968
• Eugene Kleiner -	founded Kleiner-Perkins
• Victor Grinich -	only a poor professor at UC Berkeley & Stanford

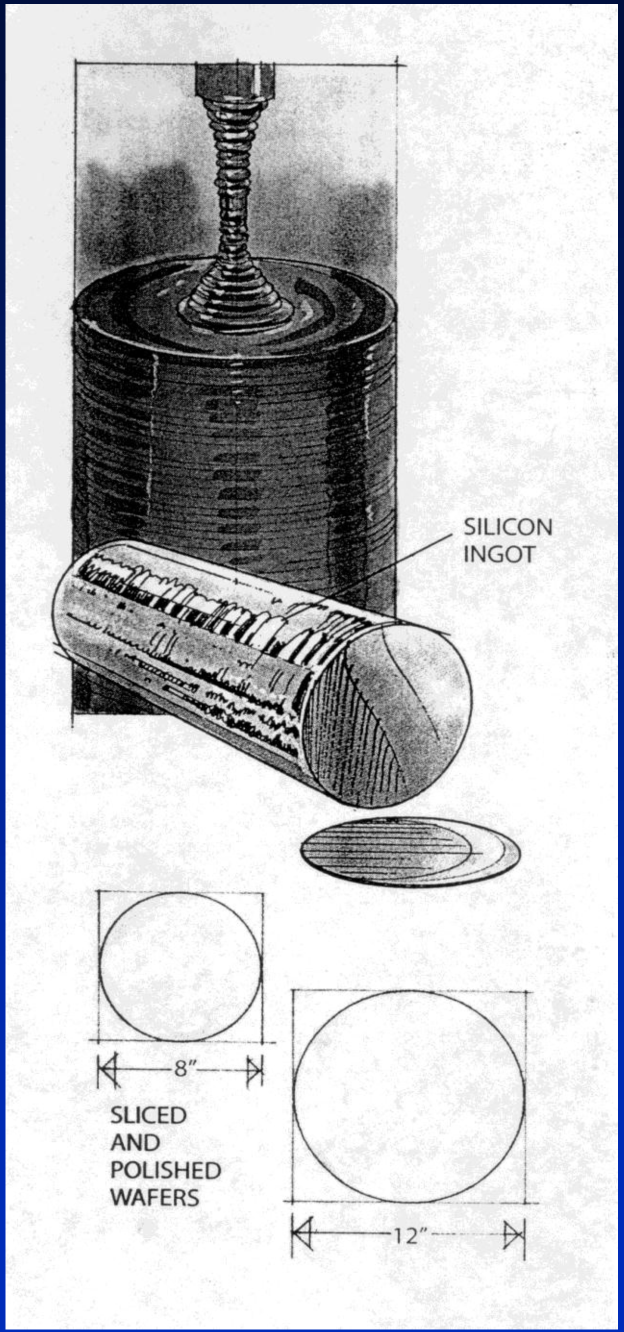
From Sand to Silicon – Manufacturing an Integrated Circuit

Scientific American: The Solid-State Century, Special Issue 1998

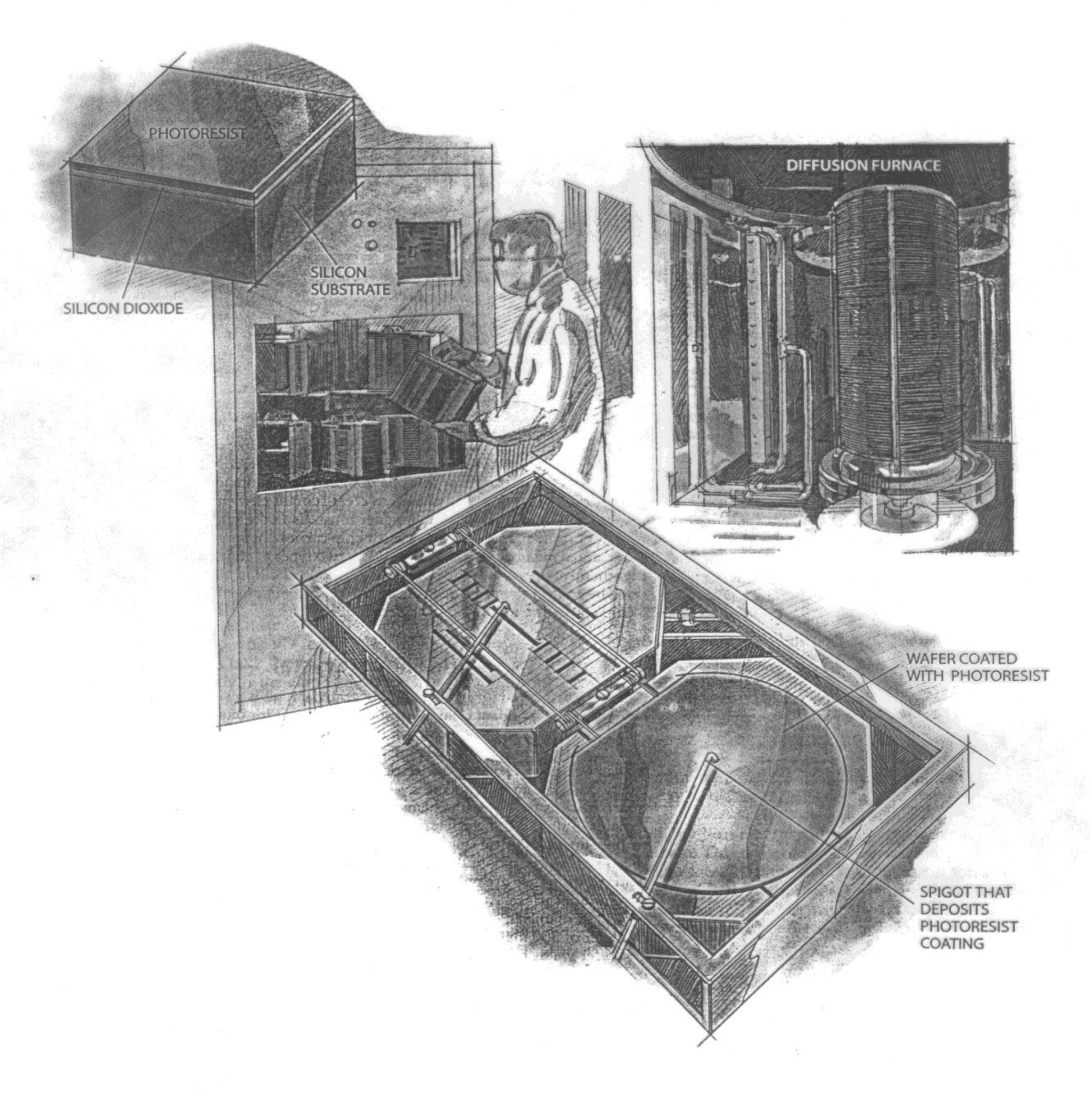
Chip Design



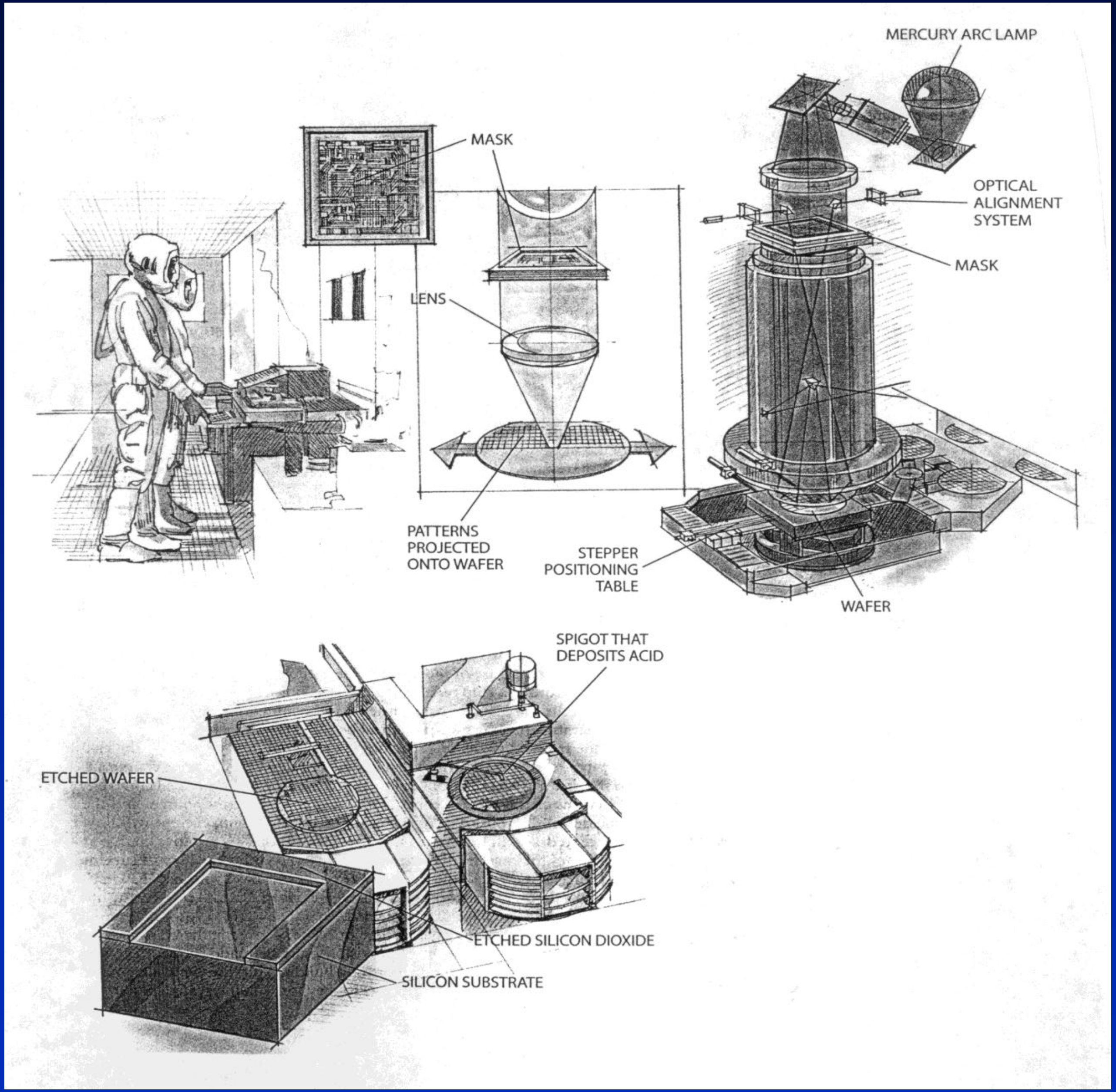
Silicon Crystal



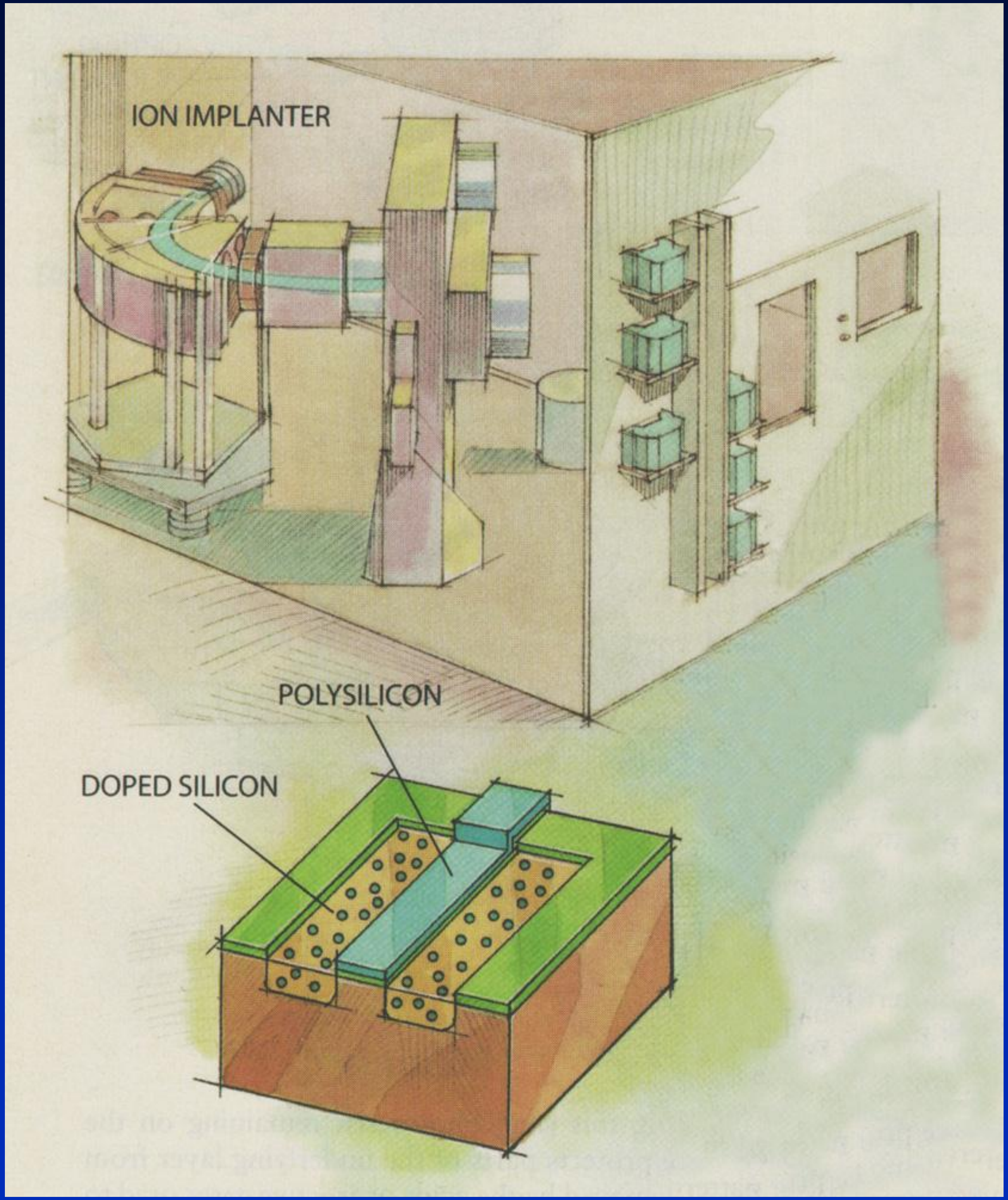
Layering


PHOTORESIST
SILICON DIOXIDE
SILICON SUBSTRATE
DIFFUSION FURNACE
WAFER COATED WITH PHOTORESIST
SPIGOT THAT DEPOSITS PHOTORESIST COATING

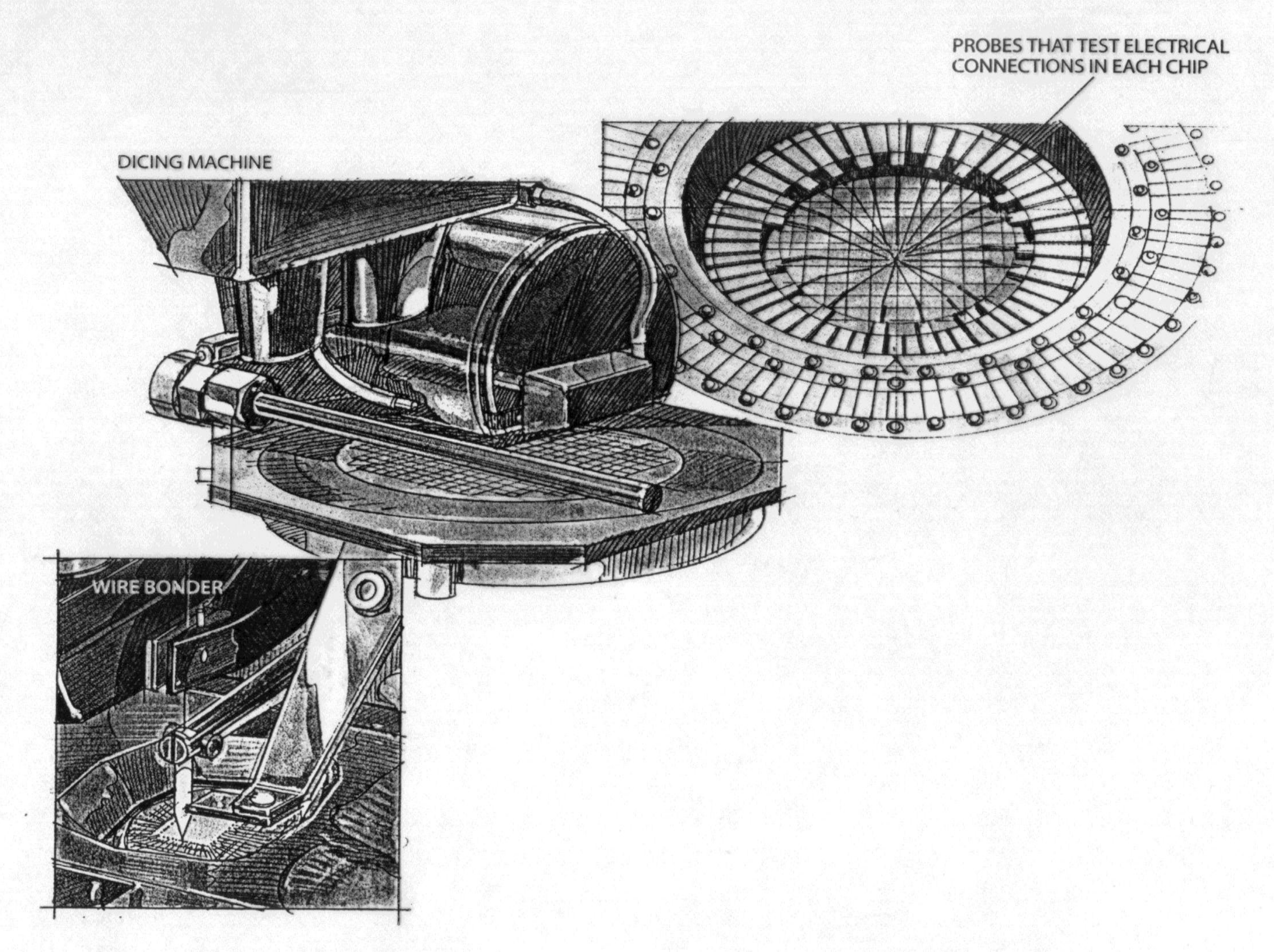
Masking & Etching



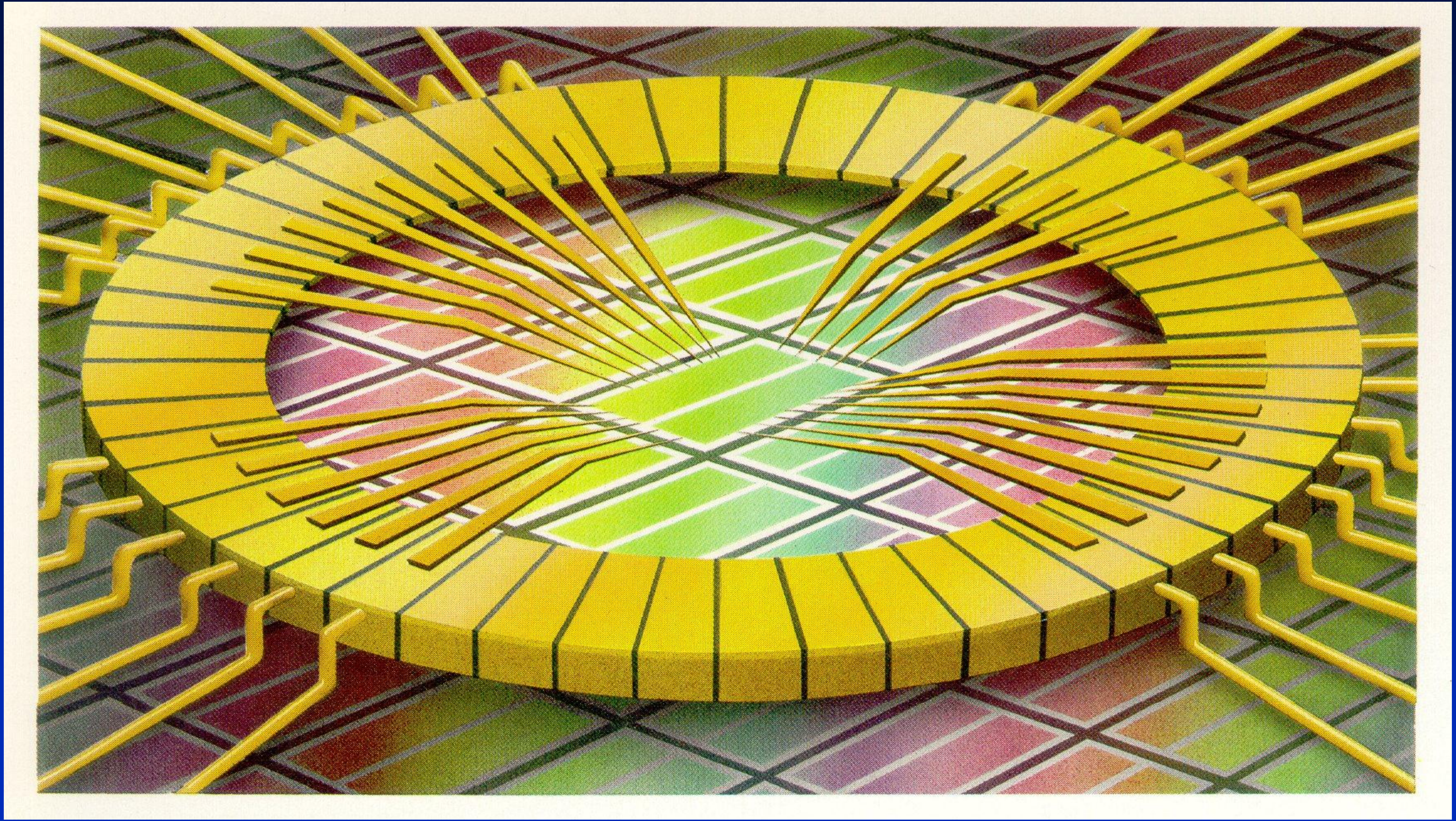
Doping



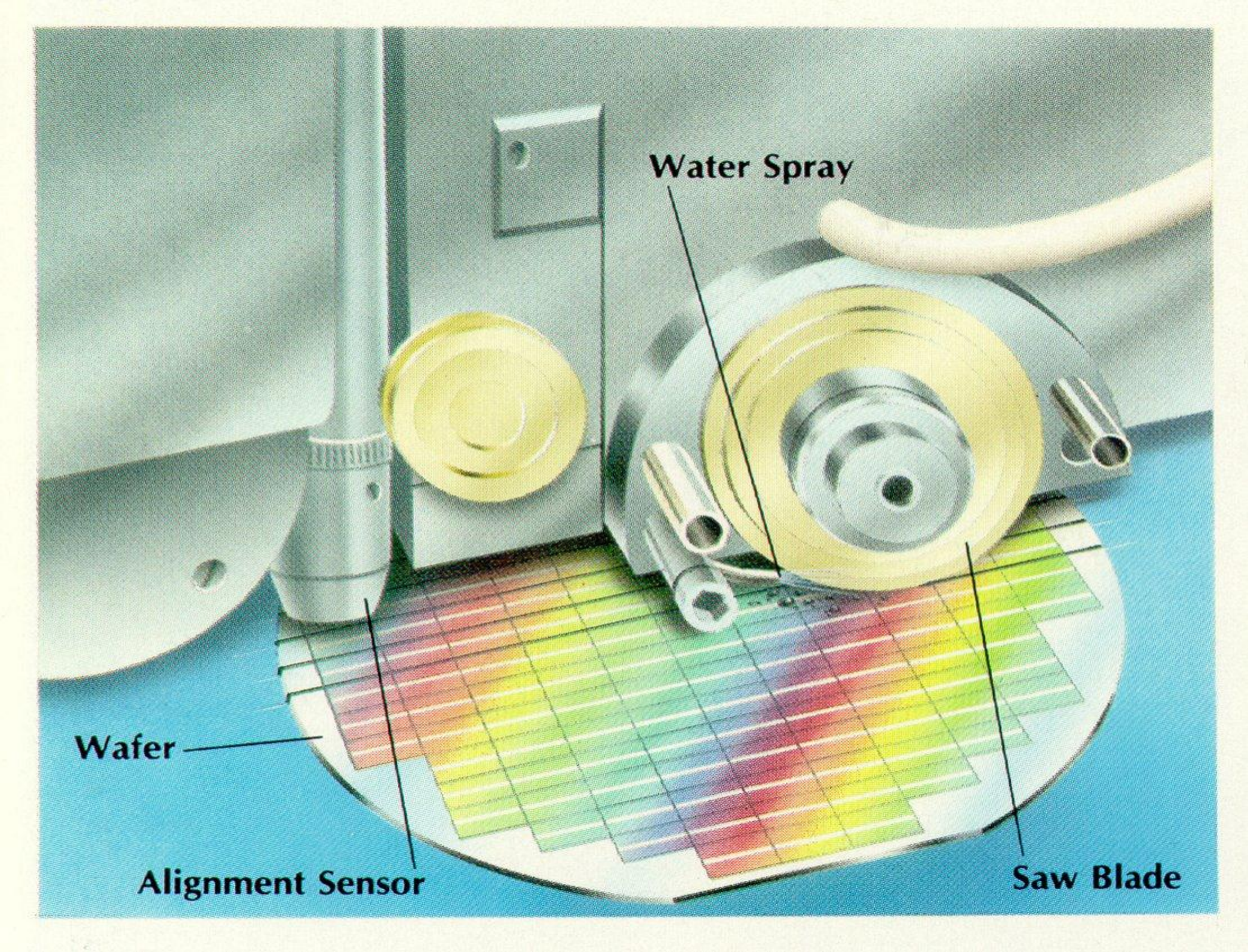
Interconnections & Dicing



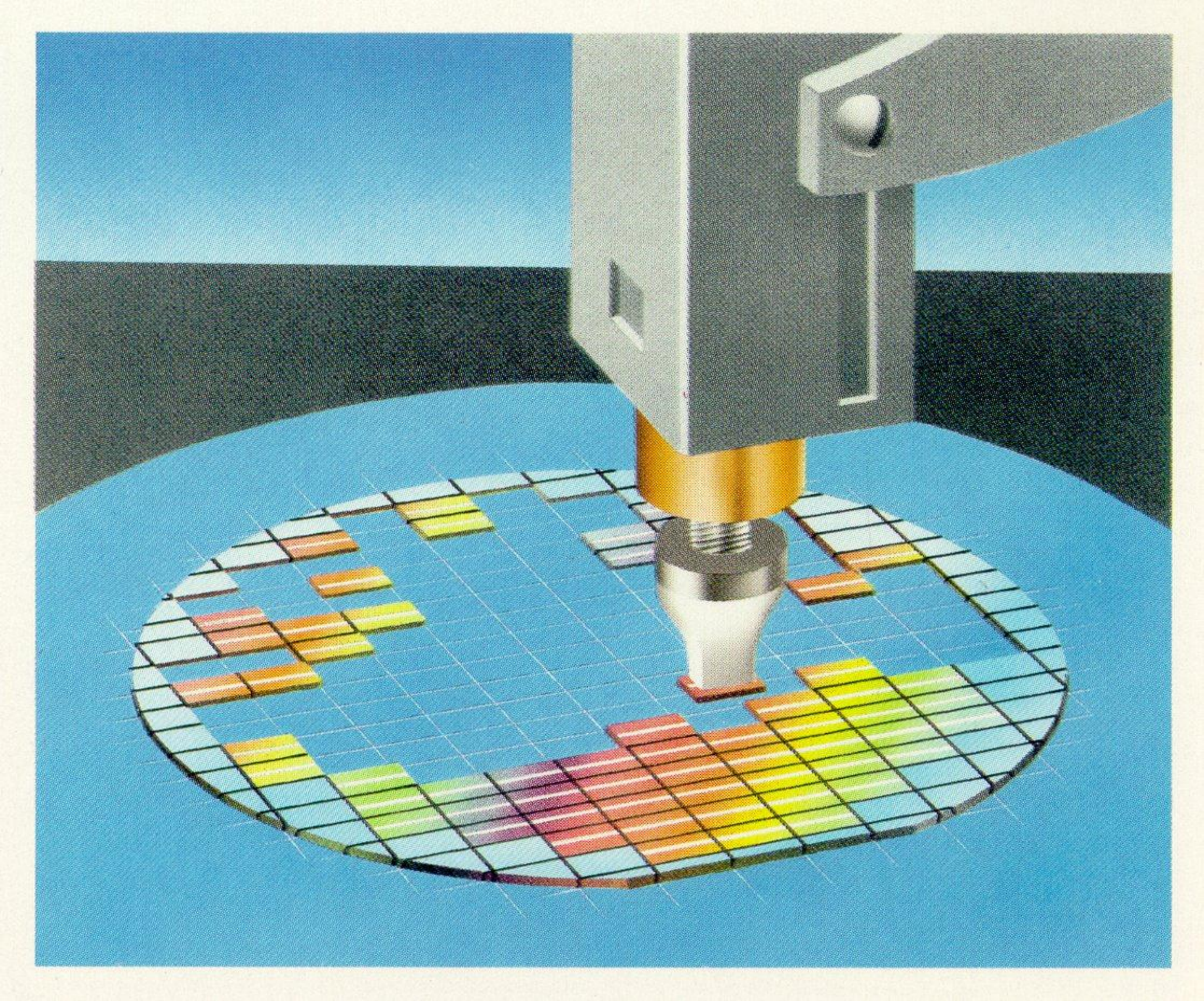
Probing Electrical Connections



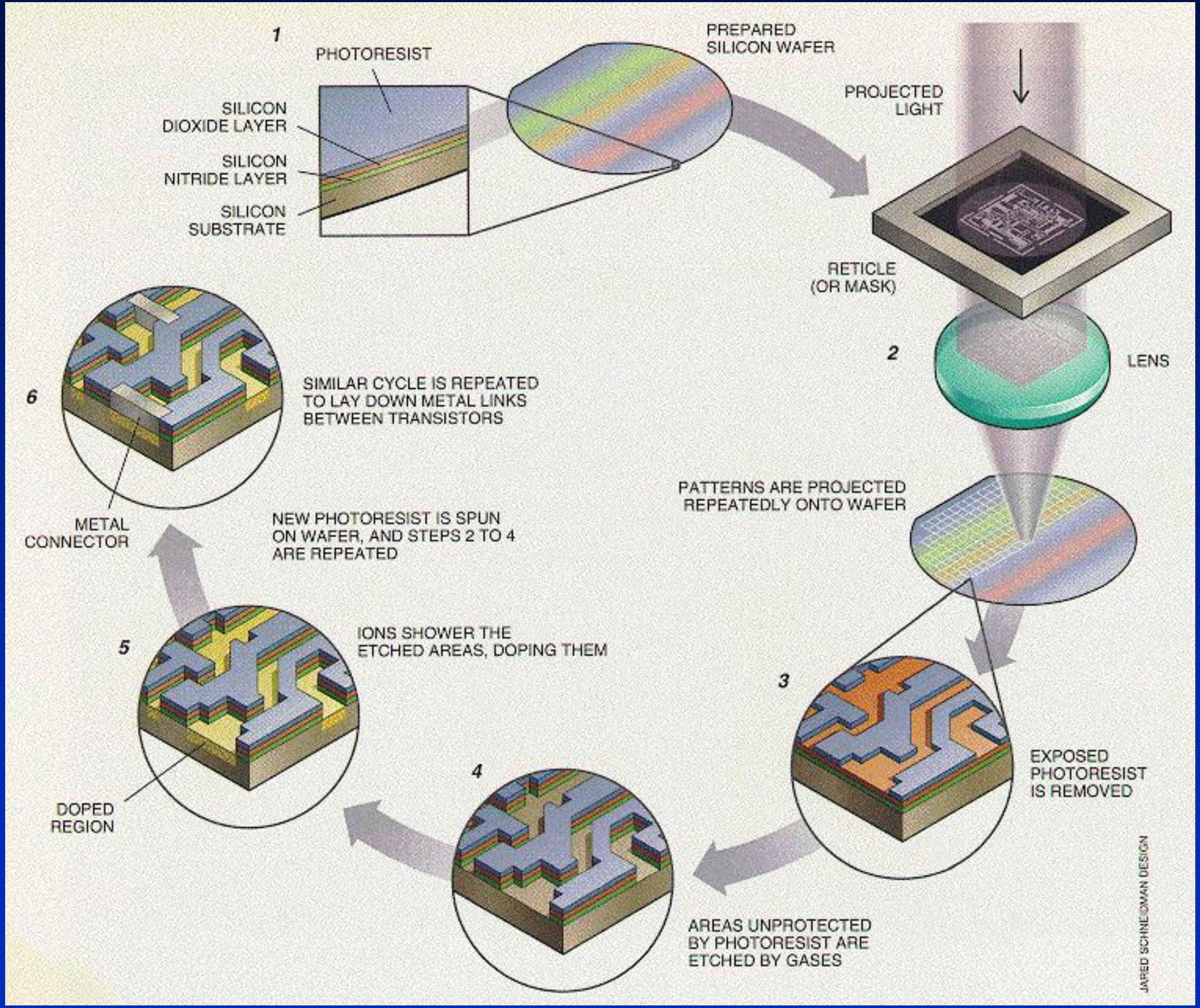
Dicing

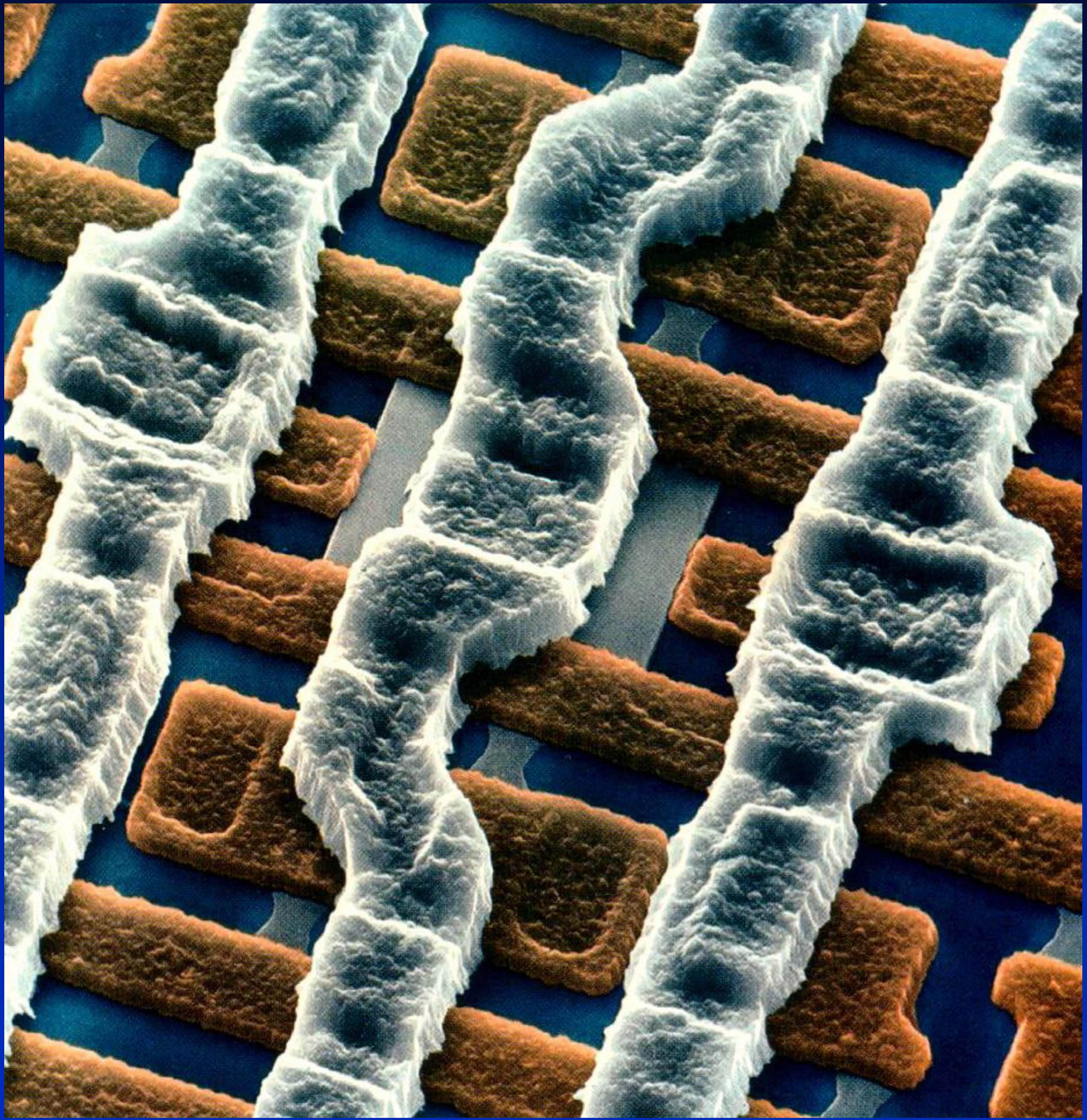


Chip Selection



Chip Fabrication


1
PHOTORESIST
SILICON DIOXIDE LAYER
SILICON NITRIDE LAYER
SILICON SUBSTRATE
PREPARED SILICON WAFER
PROJECTED LIGHT
RETICLE (OR MASK)
2
LENS
PATTERNS ARE PROJECTED REPEATEDLY ONTO WAFER
3
EXPOSED PHOTORESIST IS REMOVED
4
AREAS UNPROTECTED BY PHOTORESIST ARE ETCHED BY GASES
5
IONS SHOWER THE ETCHED AREAS, DOPING THEM
DOPED REGION
NEW PHOTORESIST IS SPUN ON WAFER, AND STEPS 2 TO 4 ARE REPEATED
6
SIMILAR CYCLE IS REPEATED TO LAY DOWN METAL LINKS BETWEEN TRANSISTORS
METAL CONNECTOR
JARED SCHNEIDMAN DESIGN

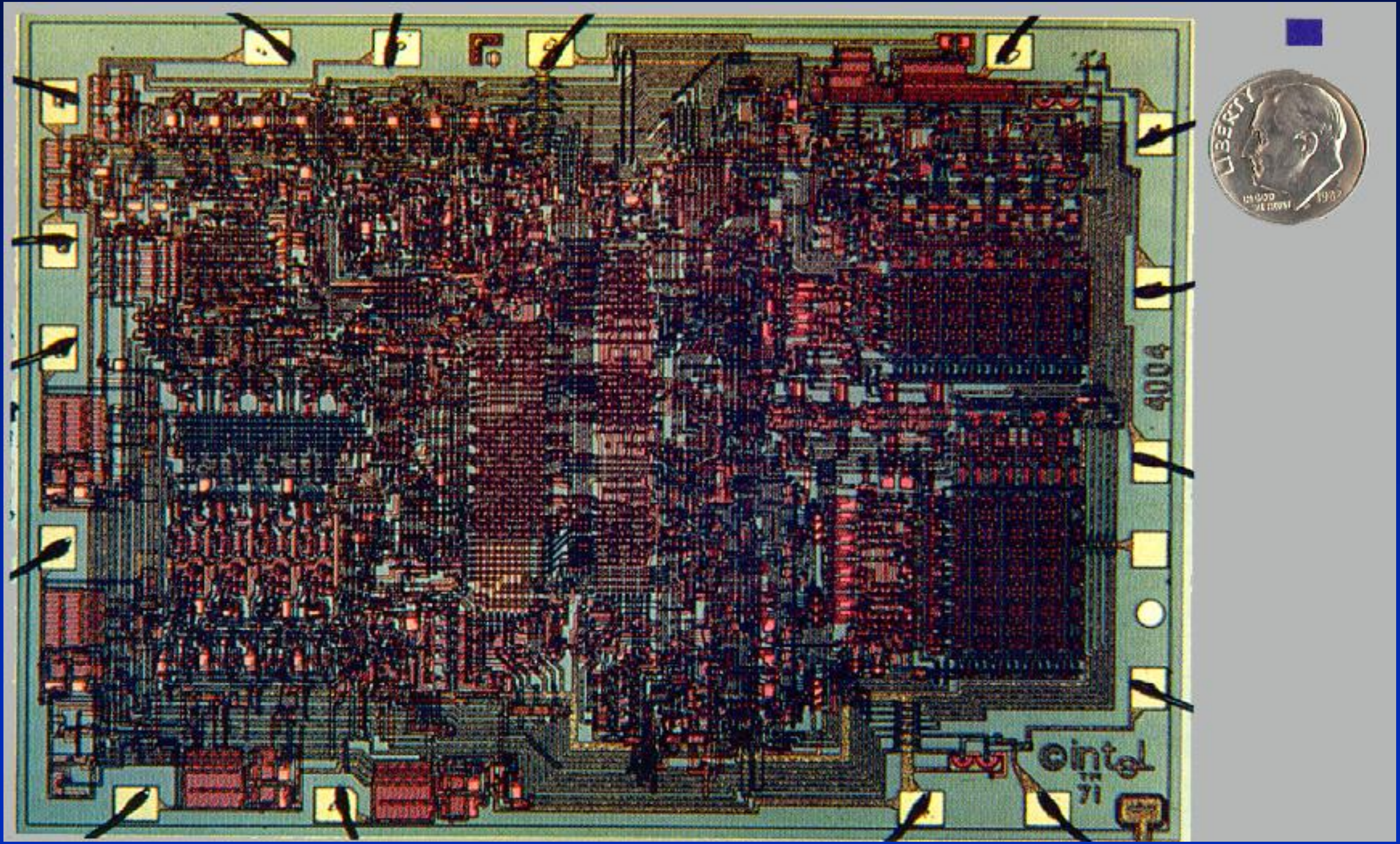


International Technology Roadmap for Semiconductors

	2001	2004	2007	2010	2013	2016
Technology (nanometers)	**130nm**	**90nm**	**65nm**	**45nm**	**32nm**	**22nm**
Functions per Chip (millions)	97	193	386	1546	3092	6184
Clock Speed (Ghz)	2.5Ghz	4.1Ghz	9.3Ghz	15Ghz	23Ghz	40Ghz
Wafer Size (millimeters)	200mm	300mm	300mm	300mm	450mm	450mm
Chip Size (mm^2)	140 mm^2	140 mm^2	140 mm^2	140 mm^2	140 mm^2	140 mm^2

Roughly 0.5 shrink every 3 years. Intel released 22 nm chips in 2013

Intel 4004 November 1971



Moore’s Law – CPU Transistor Counts

Processor	Transistor count	Date of introduction	Manufacturer	Process	Area
Intel 4004	2,300	1971	Intel	10 µm	
Intel 8008	3,500	1972	Intel	10 µm	
Intel 8080	4,500	1974	Intel	6 µm	
Intel 8088	29,000	1979	Intel	3 µm	
Intel 80286	134,000	1982	Intel	1.5 µm	
Intel 80386	275,000	1985	Intel	1.5 µm	
Intel 80486	1,180,000	1989	Intel	1 µm	
Pentium	3,100,000	1993	Intel	0.8 µm	
AMD K5	4,300,000	1996	AMD	0.5 µm	
Pentium II	7,500,000	1997	Intel	0.35 µm	
AMD K6	8,800,000	1997	AMD	0.35 µm	
Pentium III	9,500,000	1999	Intel	0.25 µm	
AMD K6-III	21,300,000	1999	AMD	0.25 µm	
AMD K7	22,000,000	1999	AMD	0.25 µm	
Pentium 4	42,000,000	2000	Intel	180 nm	
Atom	47,000,000	2008	Intel	45 nm	
Barton	54,300,000	2003	AMD	130 nm	
AMD K8	105,900,000	2003	AMD	130 nm	
Itanium 2	220,000,000	2003	Intel	130 nm	

http://en.wikipedia.org/wiki/Transistor_count

Moore’s Law – CPU Transistor Counts

Processor	Transistor count	Date of introduction	Manufacturer	Process	Area
Core 2 Duo	291,000,000	2006	Intel	65 nm	
AMD K10	463,000,000	2007	AMD	65 nm	
AMD K10	758,000,000	2008	AMD	45 nm	
Itanium 2 with 9MB cache	592,000,000	2004	Intel	130 nm	
Core i7 (Quad)	731,000,000	2008	Intel	45 nm	263 mm^2
POWER6	789,000,000	2007	IBM	65 nm	341 mm^2
Six-Core Opteron 2400	904,000,000	2009	AMD	45 nm	
Six-Core Core i7	1,170,000,000	2010	Intel	32 nm	
Dual-Core Itanium 2	1,700,000,000	2006	Intel	90 nm	596 mm^2
Six-Core Xeon 7400	1,900,000,000	2008	Intel	45 nm	
Quad-Core Itanium Tukwila	2,000,000,000	2010	Intel	65 nm	
Six-Core Core i7 (Sandy Bridge-E)	2,270,000,000	2011	Intel	32 nm	434 mm^2
8-Core Xeon Nehalem-EX	2,300,000,000	2010	Intel	45 nm	684 mm^2
10-Core Xeon Westmere-EX	2,600,000,000	2011	Intel	32 nm	512 mm^2
Six-core zEC12	2,750,000,000	2012	IBM	32 nm	597 mm^2
8-Core Itanium Poulson	3,100,000,000	2012	Intel	32 nm	544 mm^2
15-Core Xeon Ivy Bridge-EX	4,310,000,000	2014	Intel	22nm	541 mm^2
62-Core Xeon Phi	5,000,000,000	2012	Intel	22 nm	350 mm^2
Xbox One Main SoC	5,000,000,000	2013	Microsoft	28 nm	363 mm^2
18-core Xeon Haswell-E5	5,560,000,000	2014	Intel	22 nm	661mm^2
IBM z13 Storage Controller	7,100,000,000	2015	IBM	22 nm	678 mm^2

http://en.wikipedia.org/wiki/Transistor_count

Moore’s Law – GPU Transistor Counts

Processor	Transistor count	Date of introduction	Manufacturer	Process	Area
R520	321,000,000	2005	AMD	90 nm	288 mm^2
R580	384,000,000	2006	AMD	90 nm	352 mm^2
G80	681,000,000	2006	NVIDIA	90 nm	480 mm^2
R600 Pele	700,000,000	2007	AMD	80 nm	420 mm^2
G92	754,000,000	2007	NVIDIA	65 nm	324 mm^2
RV790XT Spartan	959,000,000	2008	AMD	55 nm	282 mm^2
GT200 Tesla	1,400,000,000	2008	NVIDIA	65 nm	576 mm^2
Cypress RV870	2,154,000,000	2009	AMD	40 nm	334 mm^2
Cayman RV970	2,640,000,000	2010	AMD	40 nm	389 mm^2
GF100 Fermi	3,200,000,000	Mar 2010	NVIDIA	40 nm	526 mm^2
GF110 Fermi	3,000,000,000	Nov 2010	NVIDIA	40 nm	520 mm^2
GK104 Kepler	3,540,000,000	2012	NVIDIA	28 nm	294 mm^2
Tahiti RV1070	4,312,711,873	2011	AMD	28 nm	365 mm^2
GK110 Kepler	7,080,000,000	2012	NVIDIA	28 nm	561 mm^2
RV1090 Hawaii	6,300,000,000	2013	AMD	28 nm	438 mm^2
GM204 Maxwell	5,200,000,000	2014	NVIDIA	28 nm	398 mm^2
GM200 Maxwell	8,100,000,000	2015	NVIDIA	28 nm	601 mm^2
Fiji	8,900,000,000	2015	AMD	28 nm	596 mm^2

http://en.wikipedia.org/wiki/Transistor_count

2007

Paul S. Otellini
Intel Corporation’s fifth CEO

Forget angels on pins: Paul Otellini gets 2,500 Silverthorne chips on a silicon wafer.

Why are we continuing to strive for smaller and smaller technology?

- More transistors/chip → increased functionality and performance

- Higher speeds → partially depends on how close together the components are placed

- Cheaper – more chips/wafer, greater yields

Yield Ratio

$$yield = \frac{n_w}{n_t}$$

$$n_w = yield \bullet n_t$$

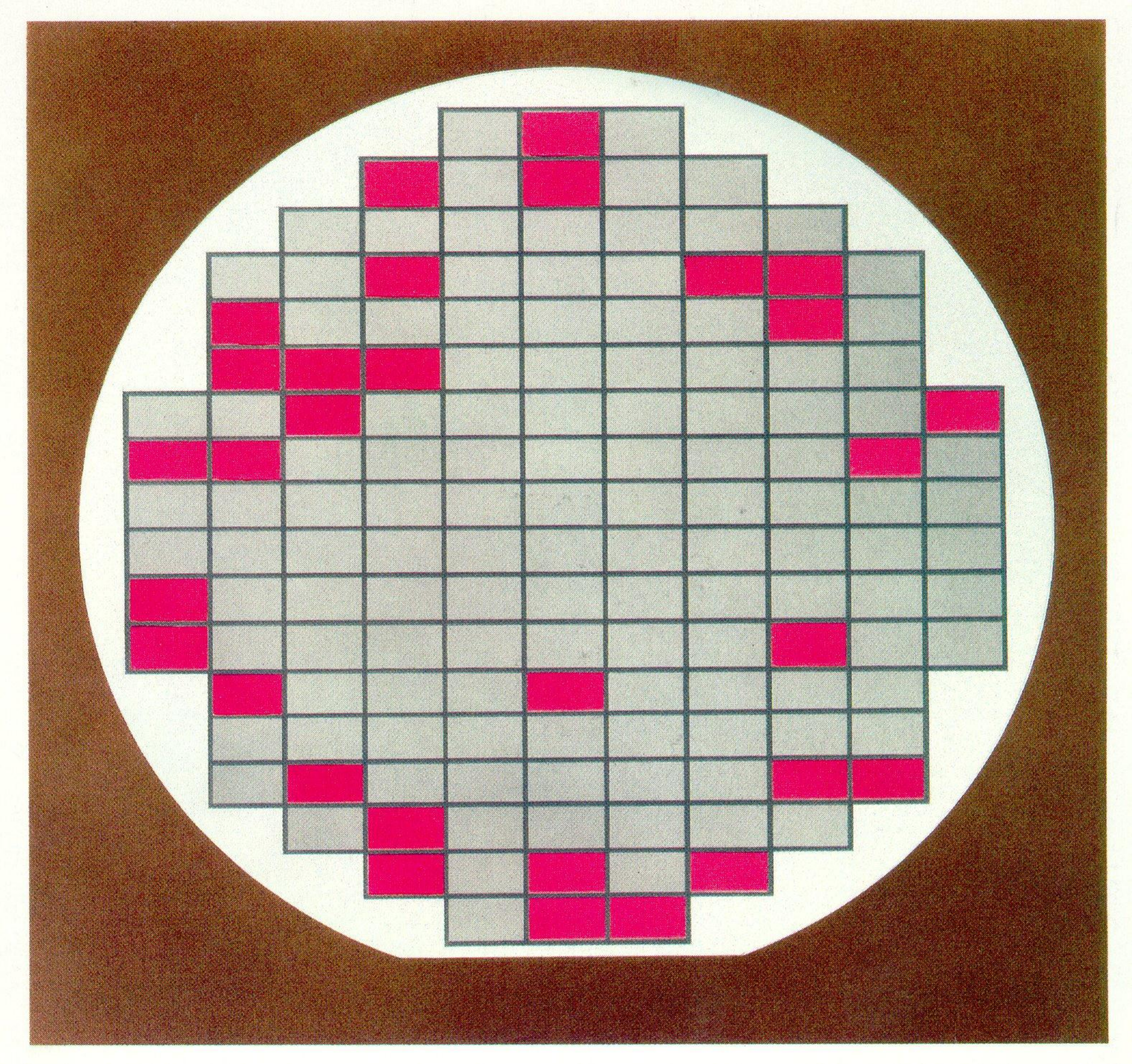
$$n_w = number\ of\ \ working\ chips/wafer$$

$$n_t = total\ number\ of\ \ chips/wafer$$

$$Old\ \ fab\ lines, yield \rightarrow > 90\%$$

$$New\ \ fab\ lines, yield \rightarrow < 40\%$$

Yield per Wafer

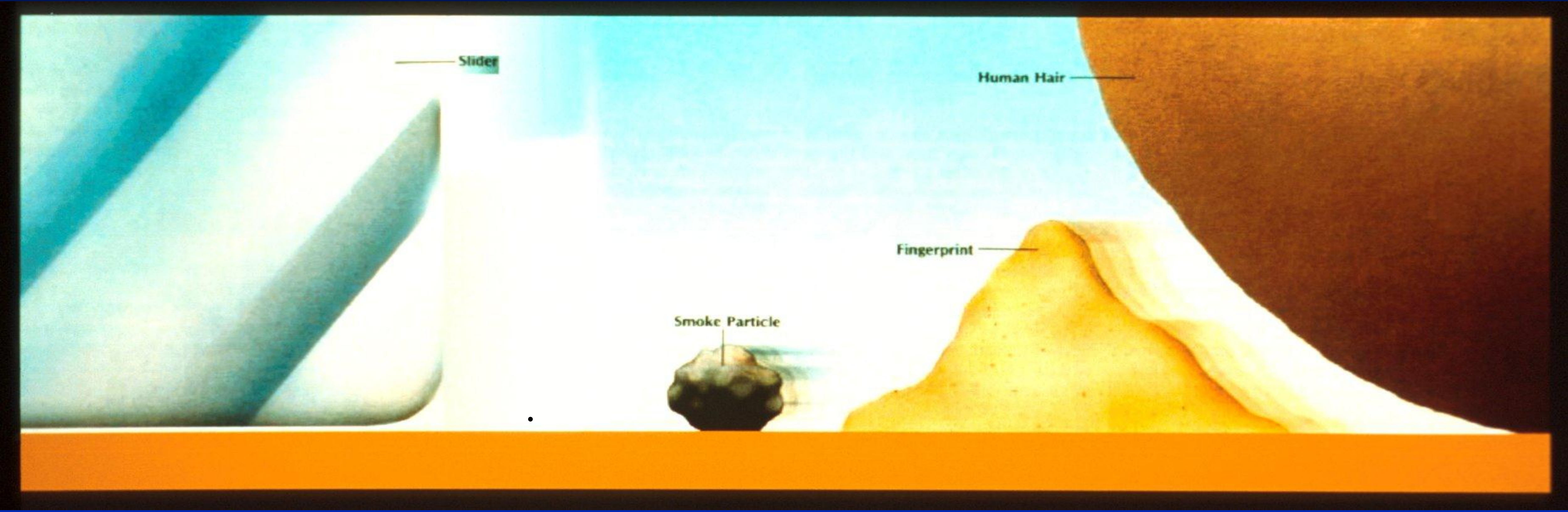


Yield Ratio

Number of defects/unit area depends on the process

$$\therefore Yield \approx \frac{1}{Chip\ Area}$$

Total chips (n_t) for a given wafer size is also inversely proportional to the chip area


Slider
Human Hair
Fingerprint
Smoke Particle

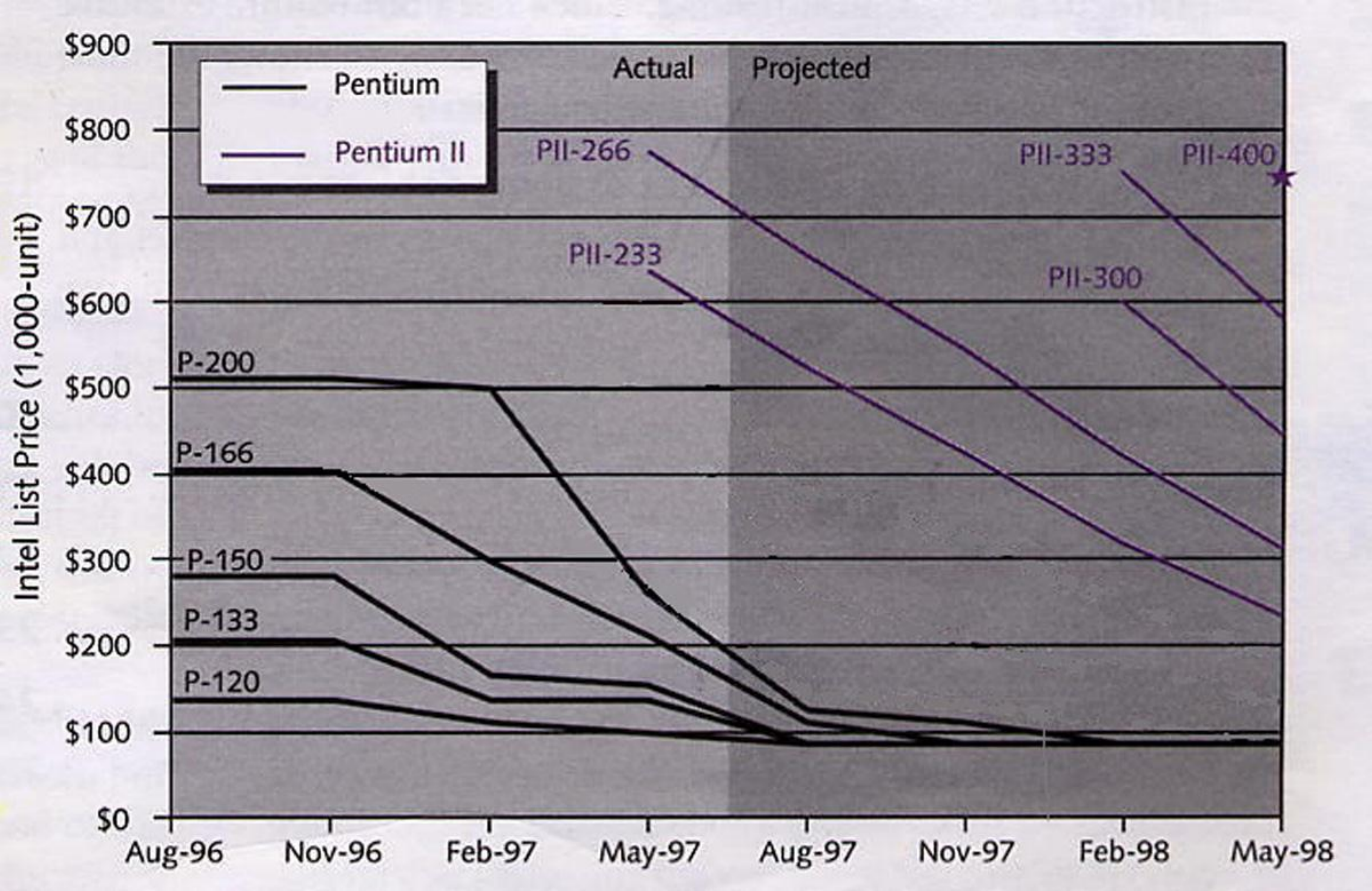
Why does the shrinking technology make the cost of manufacturing cheaper per component?

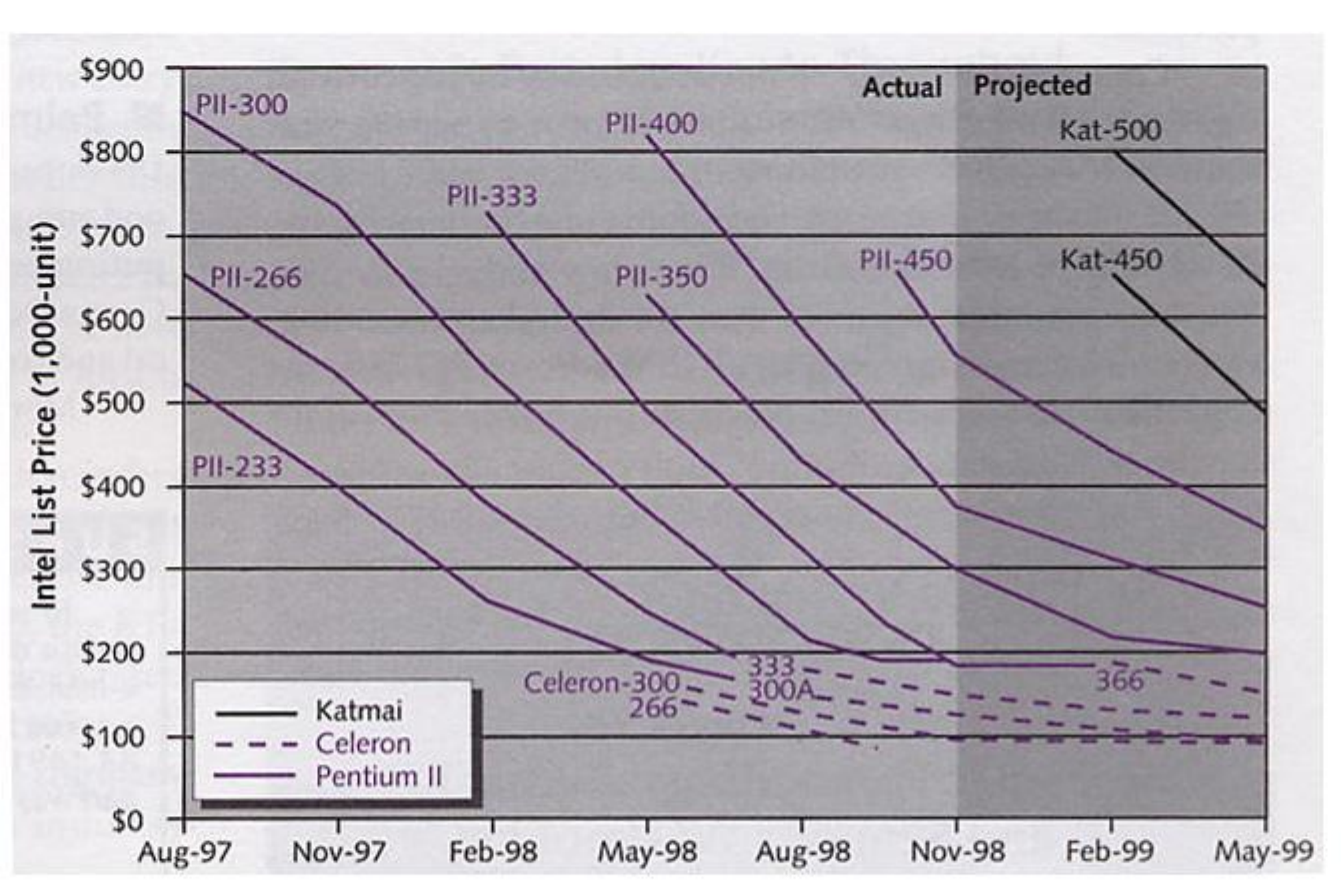
Example:

For a 10% shrink in feature size :

$$n_{w_{new}} = n_{w_{old}} \underset{\substack{\uparrow \\ \text{New yield}}}{\left(\frac{1}{.9}\right)^2} \underset{\substack{\uparrow \\ \text{New } n_t}}{\left(\frac{1}{.9}\right)^2}$$

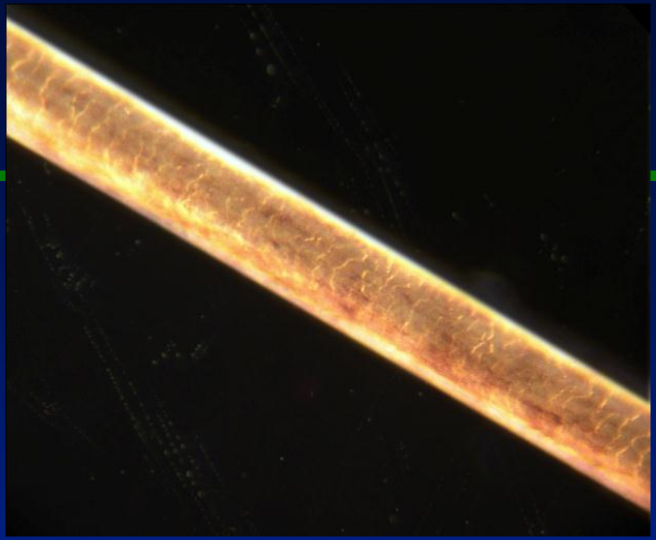
$$n_{w_{new}} = 1.52 n_{w_{old}}$$


Intel List Price (1,000-unit)
$900
$800
$700
$600
$500
$400
$300
$200
$100
$0
Aug-96
Nov-96
Feb-97
May-97
Aug-97
Nov-97
Feb-98
May-98
Actual
Projected
Pentium
Pentium II
P-200
P-166
P-150
P-133
P-120
PII-266
PII-233
PII-333
PII-400
PII-300



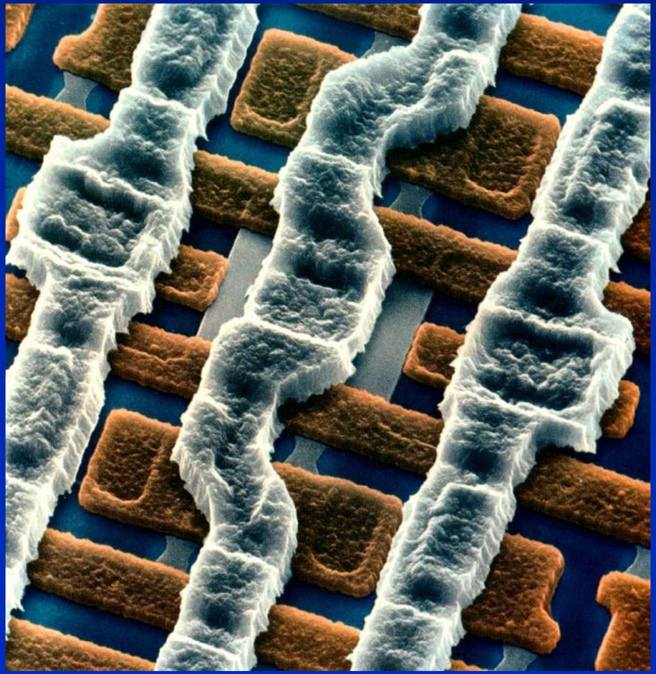
Intel List Price (1,000-unit)
$900
$800
$700
$600
$500
$400
$300
$200
$100
$0
Aug-97
Nov-97
Feb-98
May-98
Aug-98
Nov-98
Feb-99
May-99
Actual
Projected
PII-300
PII-400
Kat-500
PII-333
PII-266
PII-350
PII-450
Kat-450
PII-233
333
Celeron-300
300A
266
366
Katmai
Celeron
Pentium II

Can this Shrinking Technology Continue?



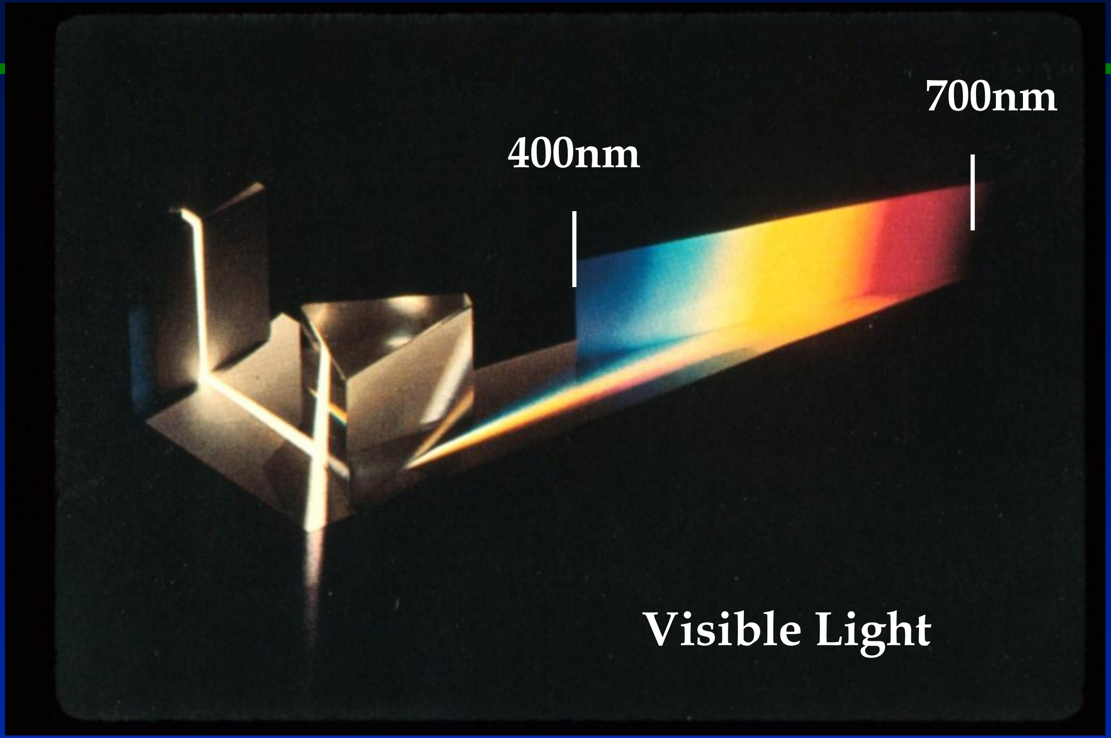
Human Hair

50µm
(50,000 nm)



Integrated Chip

32nm

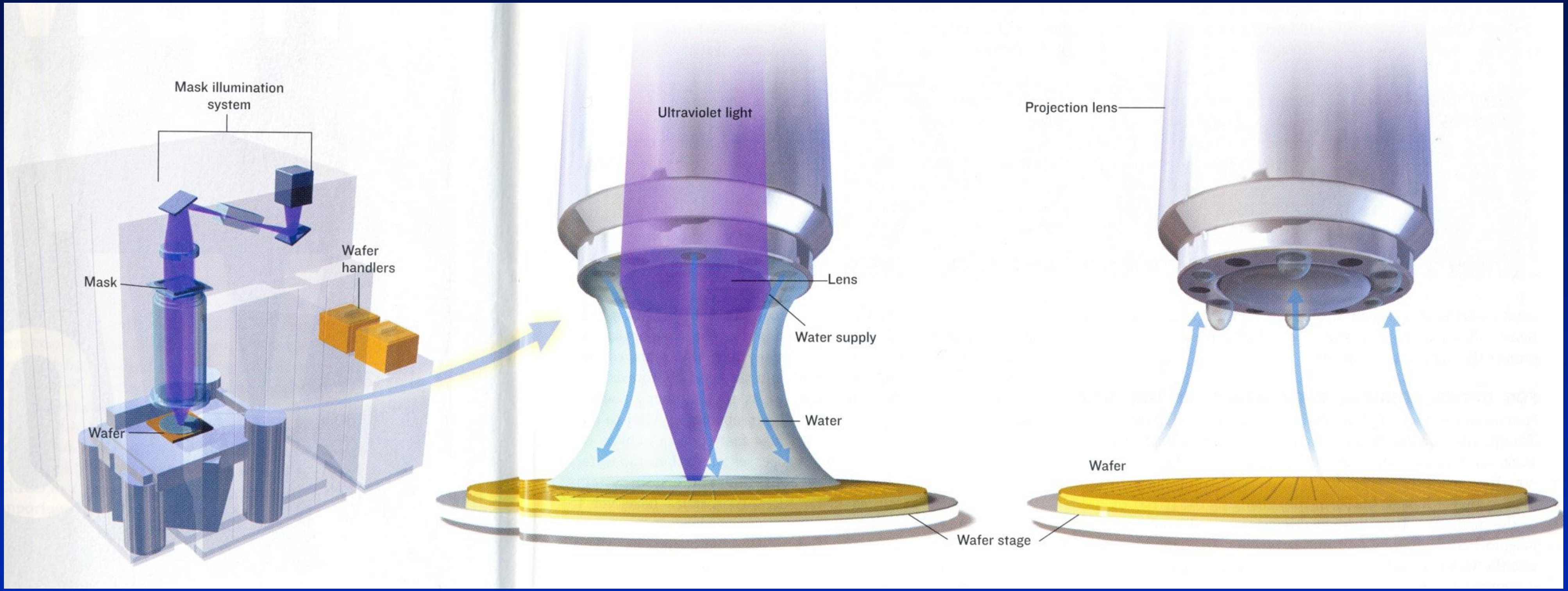


Visible Light

Keeping Up with Moore’s Law

Remarkably, Moore's Law-the number of transistors that can fit on a microchip will double every 18-24 months-has held true for many years. Keeping up with this famous prediction by Intel founder Gordon Moore is getting harder.

Getting Wafers Wet



By adding a thin layer of water between the projection lens and the wafer, the immersion system can create features 30 percent smaller.

Photolithography

- Defining the smallest components requires short wavelengths of light.
- Currently, most fabrication processors use extreme ultra-violet light at 193nm.
- Can pass the light through water. The water slows the light (less velocity) shrinking its wavelength. It is estimated that this technique will meet demands for 7 more years.
- On February 20, 2006 IBM Almaden & JSR Micro demonstrated a system using an “unidentified” light slowing liquid yielding patterns 29.9nm wide.

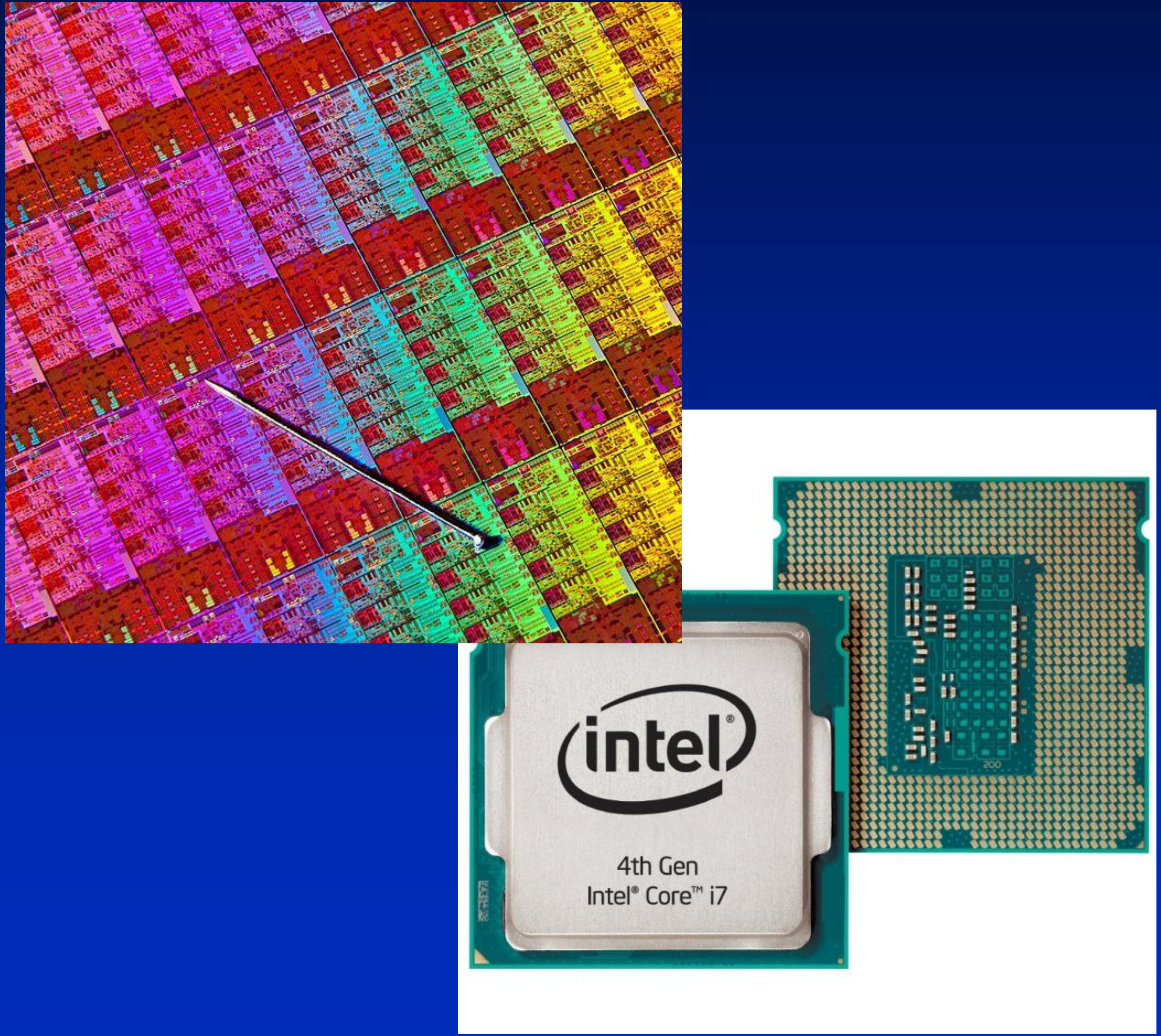
Science News, March 2, 2006

The HIGH-k SOLUTION

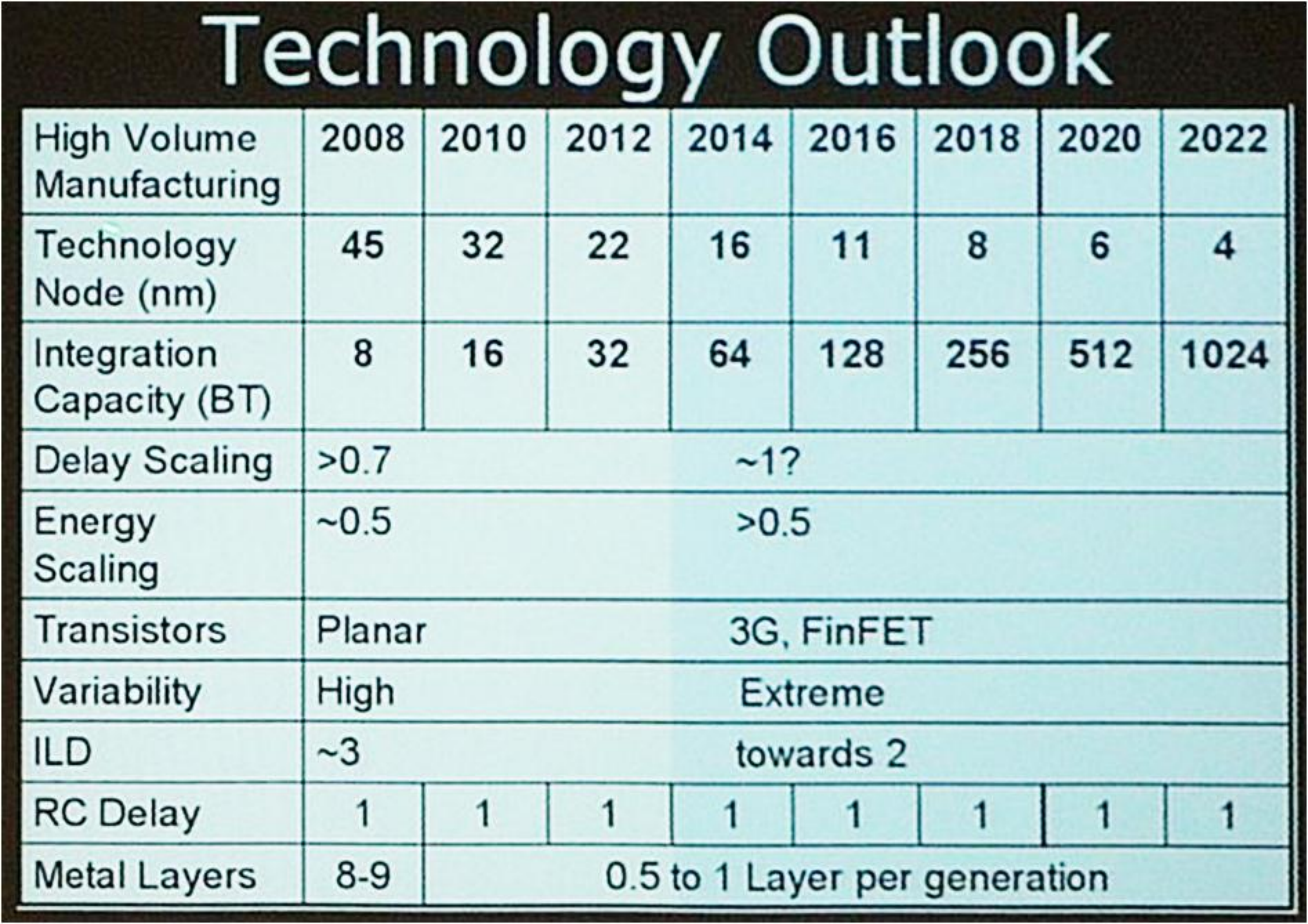
By Mark T. Bohr, Robert S. Chau, Tahir Ghani
& Kaizad Mistry – October 2007 IEEE Spectrum

In 2007 new 45nm Microprocessors were the result of the first big material redesign in CMOS transistors since the late 1960s

Intel 4th Generation i7 Chip (Haswell) - 2013



- 4 core/8 threads for desktop and mobile solutions
- 22nm Hi-K+ process
- 3D tri-gate transistors



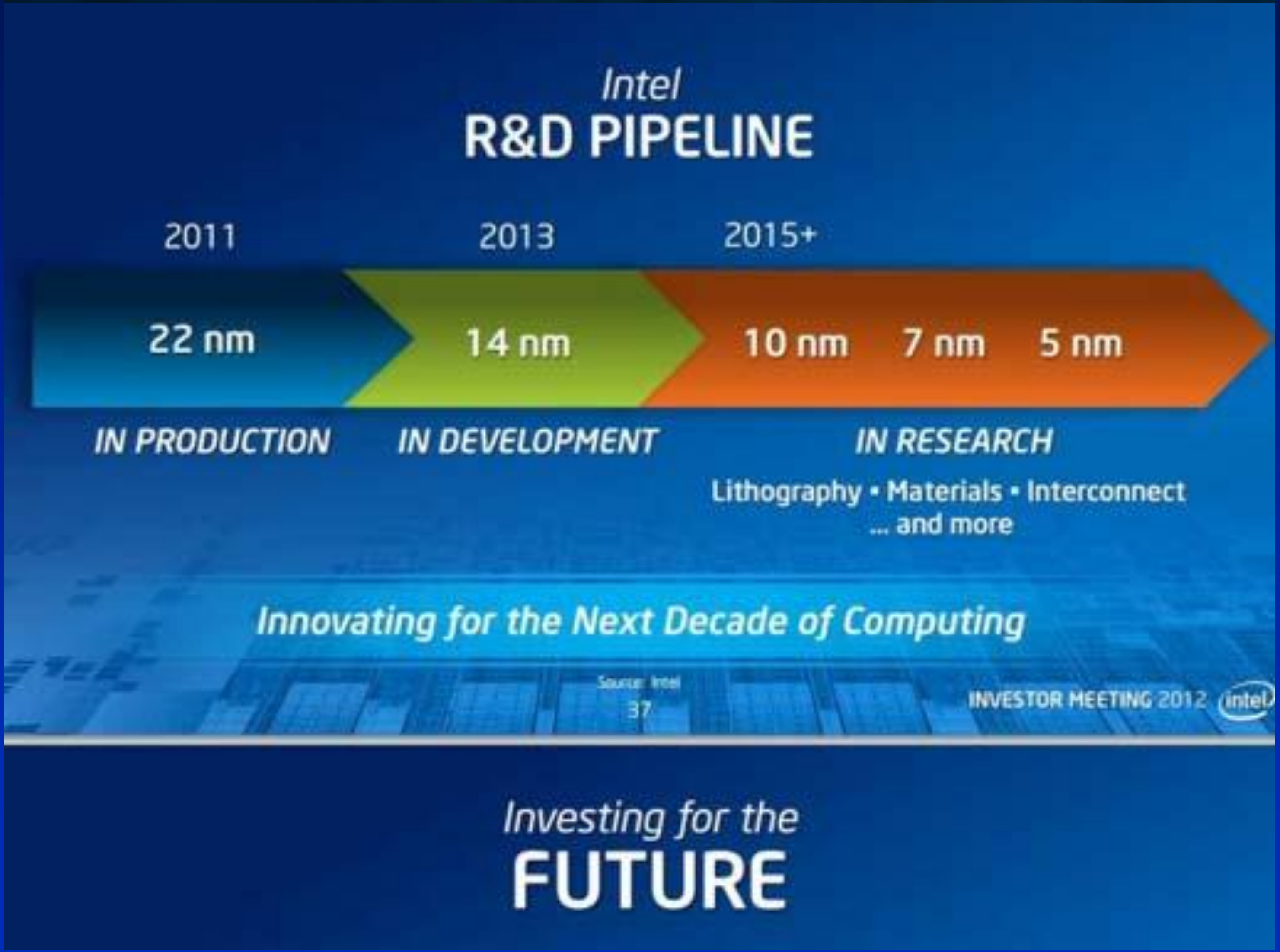
Technology Outlook

High Volume Manufacturing	2008	2010	2012	2014	2016	2018	2020	2022
Technology Node (nm)	45	32	22	16	11	8	6	4
Integration Capacity (BT)	8	16	32	64	128	256	512	1024
Delay Scaling	>0.7			~1?				
Energy Scaling	~0.5			>0.5				
Transistors	Planar			3G, FinFET				
Variability	High			Extreme				
ILD	~3			towards 2				
RC Delay	1	1	1	1	1	1	1	1
Metal Layers	8-9	0.5 to 1 Layer per generation						

10 Nanometer Technology

- Nov. 15, 2012, Samsung unveiled a 64 gigabyte (GB) multimedia card (eMMC) based on 10 nm technology.
- April 11, 2013, Samsung announced it was mass-producing High-Performance 128-gigabit NAND Flash Memory with 10 nm and 20 nm technology.
- April 2015, TSMC announced that 10 nm production would begin at the end of 2016.
- May 23rd 2015, Samsung Electronics showed off a wafer of 10nm FinFET chips.

Moore’s Law



Intel

R&D PIPELINE

2011 2013 2015+

22 nm — 14 nm — 10 nm 7 nm 5 nm

IN PRODUCTION — IN DEVELOPMENT — IN RESEARCH

Lithography • Materials • Interconnect
... and more

Innovating for the Next Decade of Computing

Source: Intel

37

INVESTOR MEETING 2012 intel

Investing for the

FUTURE

Factors Contributing to Advancing Microprocessor Performance

- Shrinking Component Size
- Increasing Speed
- Reducing Circuit Resistance
- New Materials

Factors Contributing to Advancing Microprocessor Performance

- RISC vs. CISC
- VLIW
- Multi-level Cache
- Parallelism & Pipelining

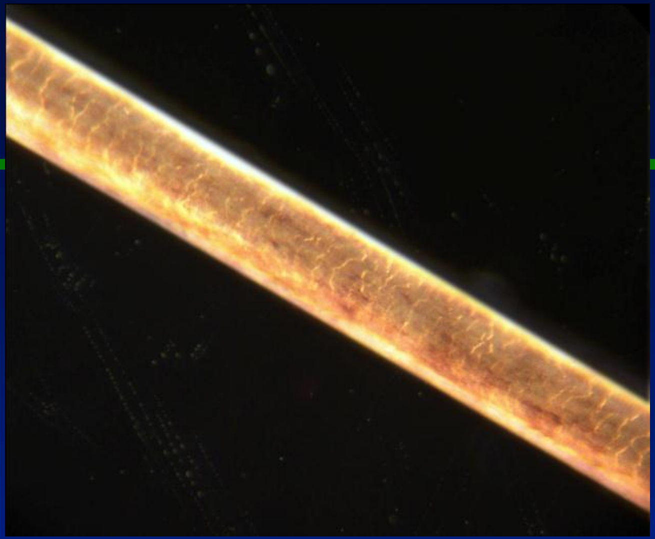
Factors Contributing to Advancing Microprocessor Performance

- RISC vs. CISC
- VLIW
- Multi-level Cache
- Parallelism & Pipelining
- Multi-core Technology

Multicore Craze

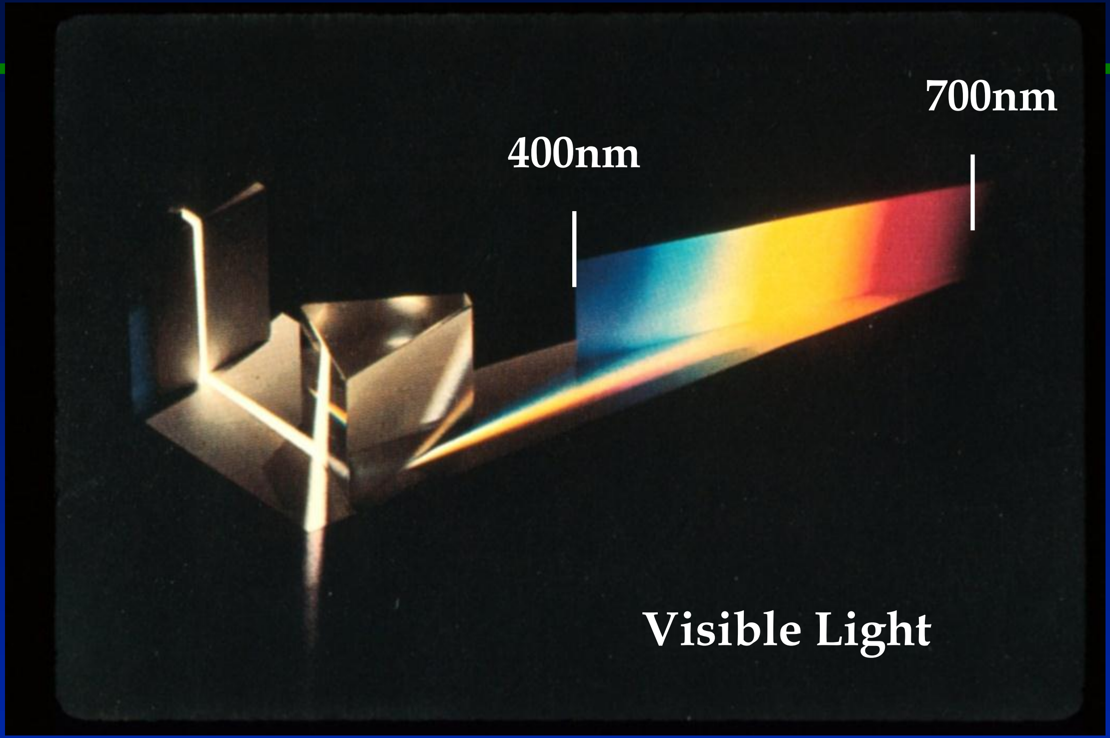
- For years, the trend was to make chips faster

  Today → 3 Ghz
- But the power required (Watts) and the heat generated is proportional to the frequency squared.
- Therefore, put more computers on the chip but run at slower speeds.

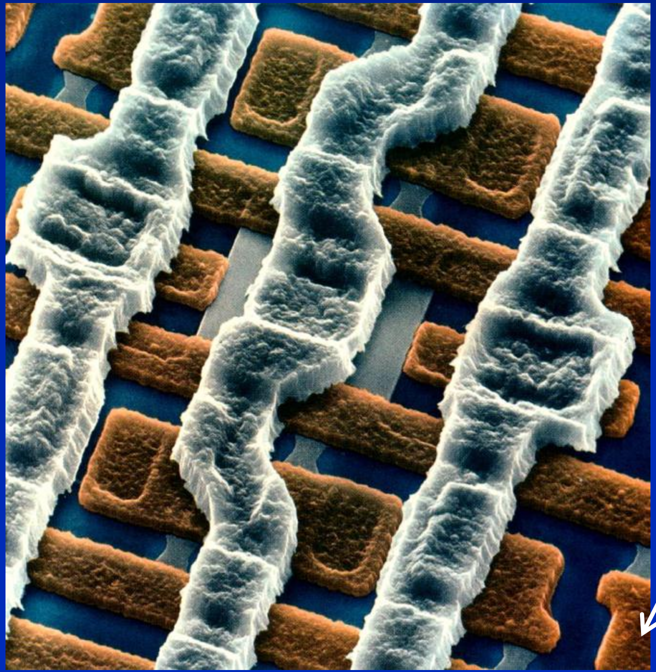


Human Hair

50µm

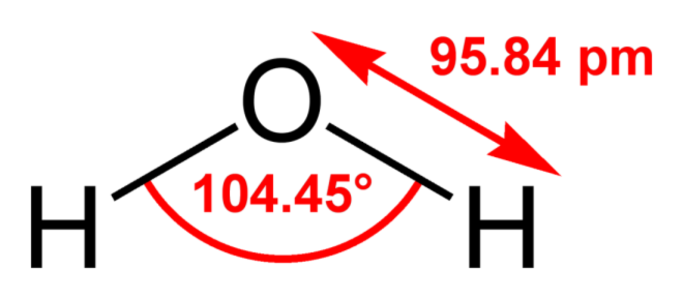


Visible Light



Integrated Chip

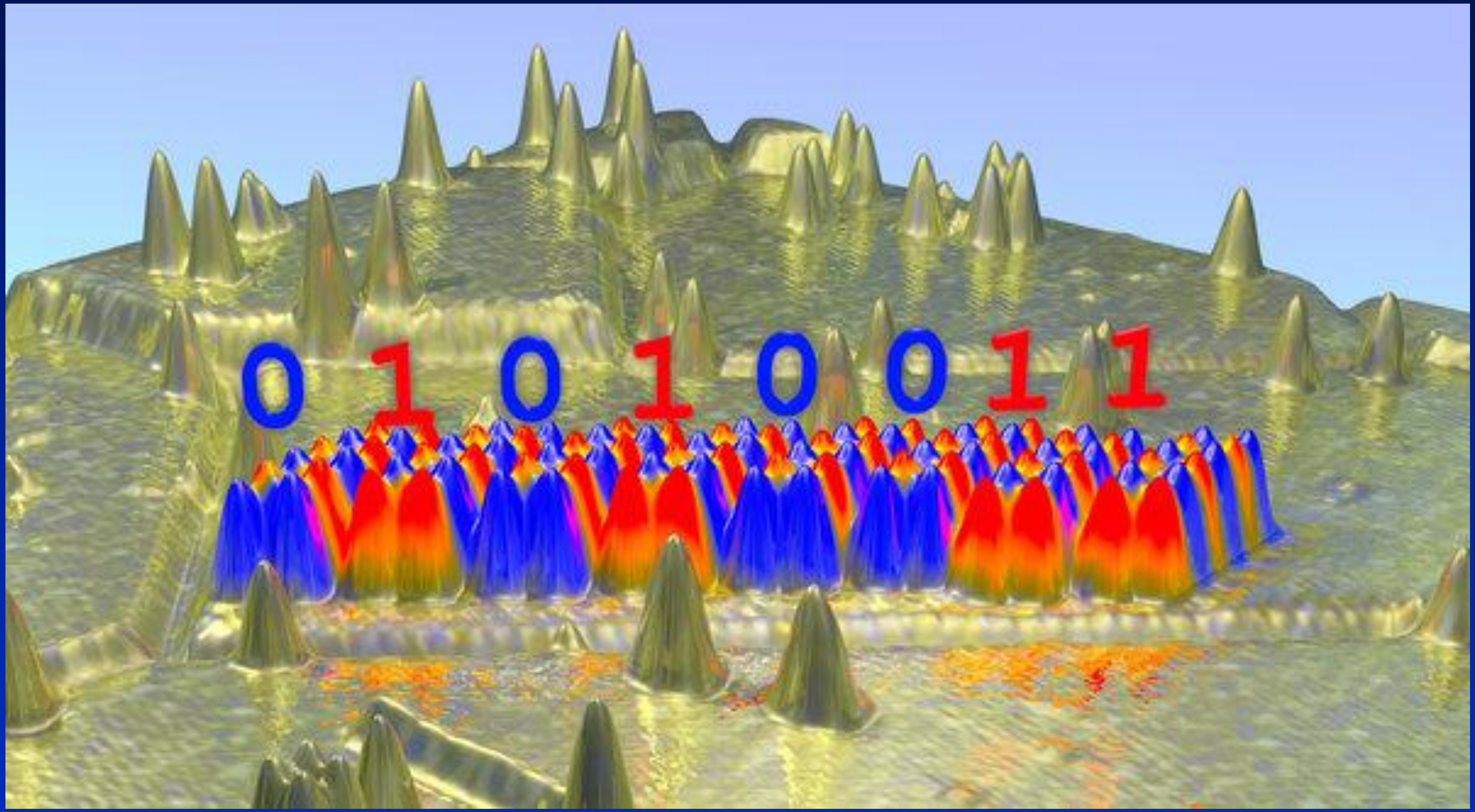
32nm



- How long can Moore’s Law continue?
- What are the limits to this integrated circuit technology?

> *“There are two constraints:*
>
> - *The finite velocity of light*
> - *The atomic nature of materials”*
>
> \- Stephen Hawking

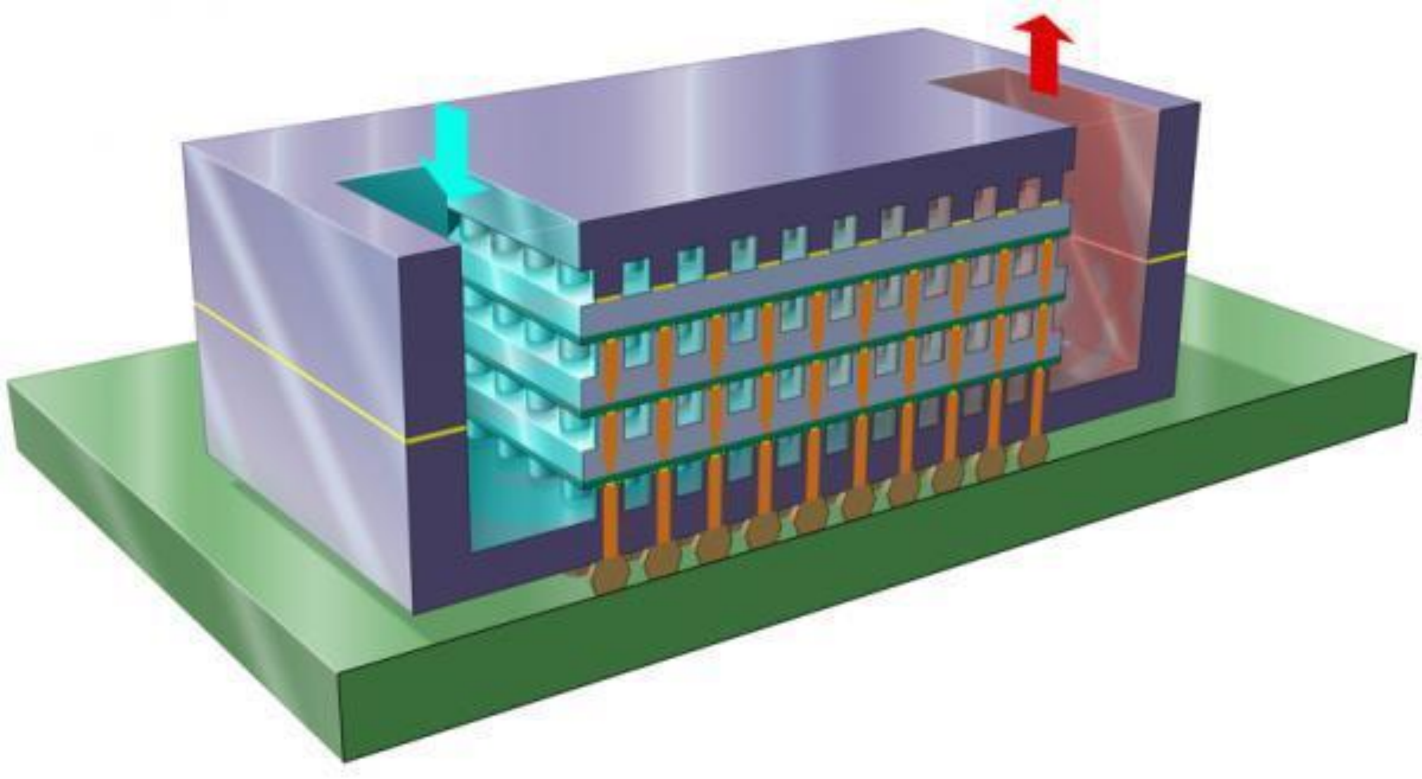
Miniaturized Data Storage at Atomic Scale



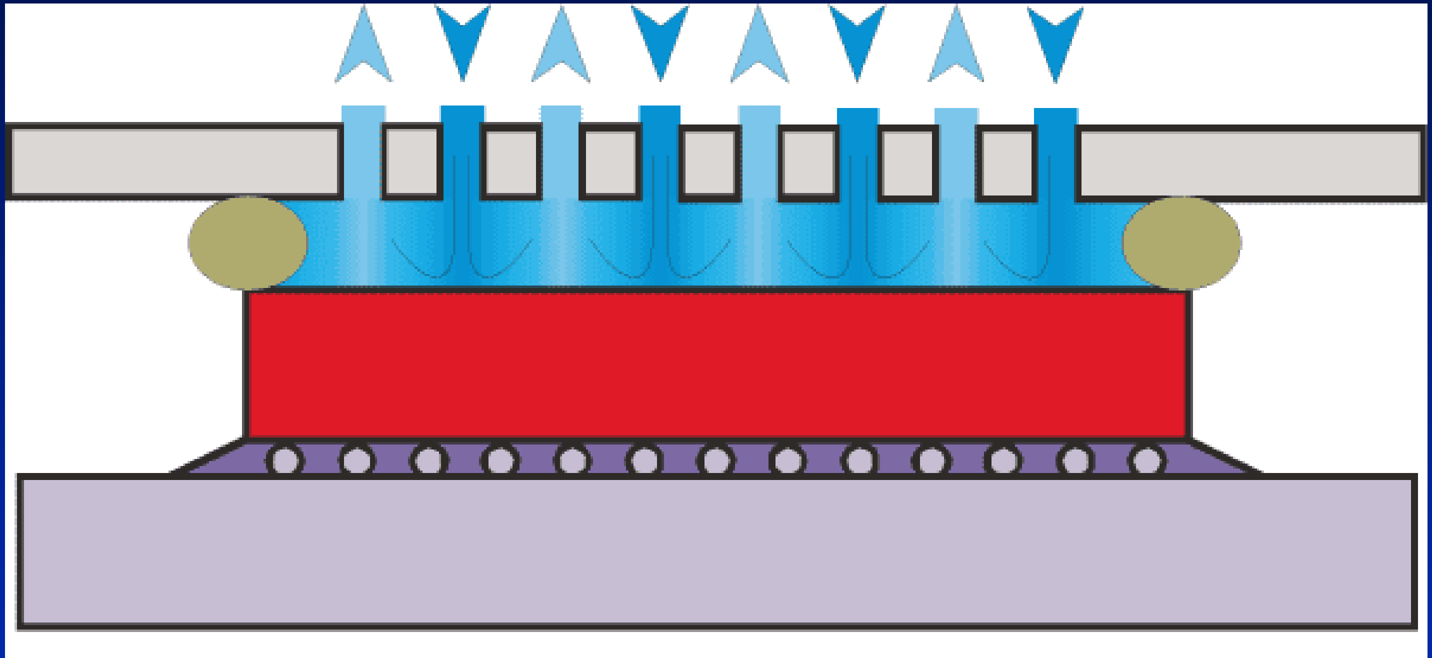
IBM researchers have stored and retrieved digital data from an array of just 12 atoms

New York Times, 1/12/12

IBM’s Chip Stacking Technology

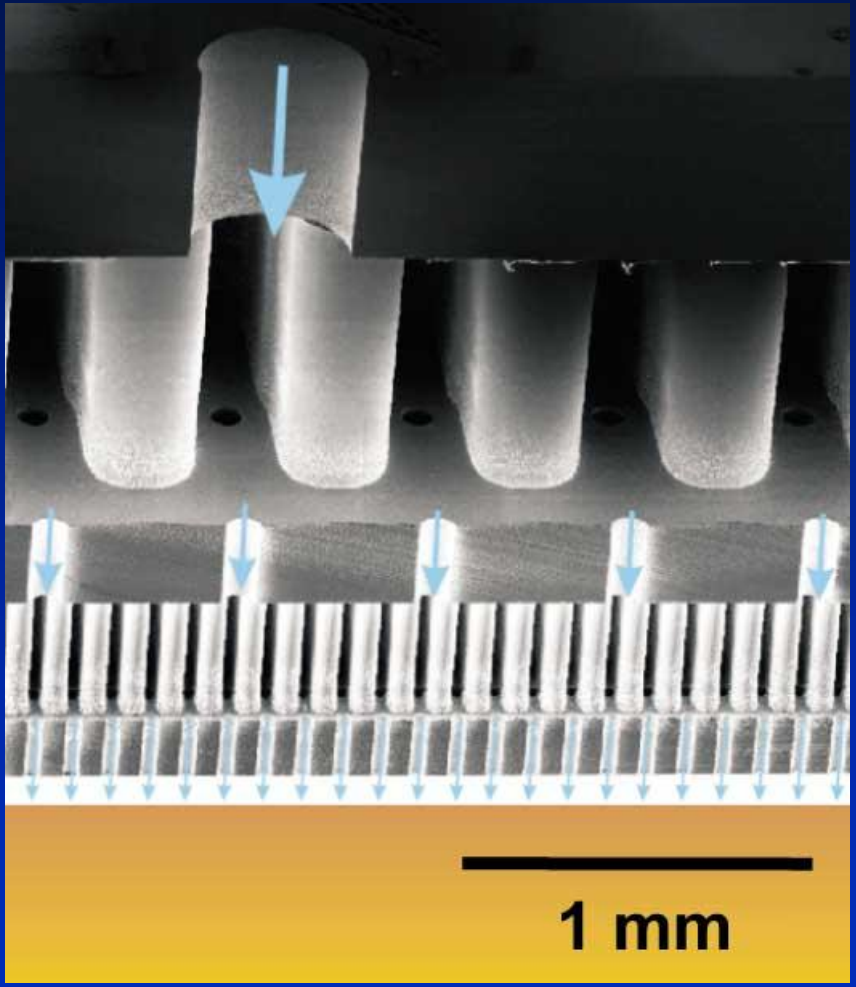


Single-phase, miniaturized convective cooling



Distributed return architecture with cross section showing inlet jets with neighboring drainage holes.

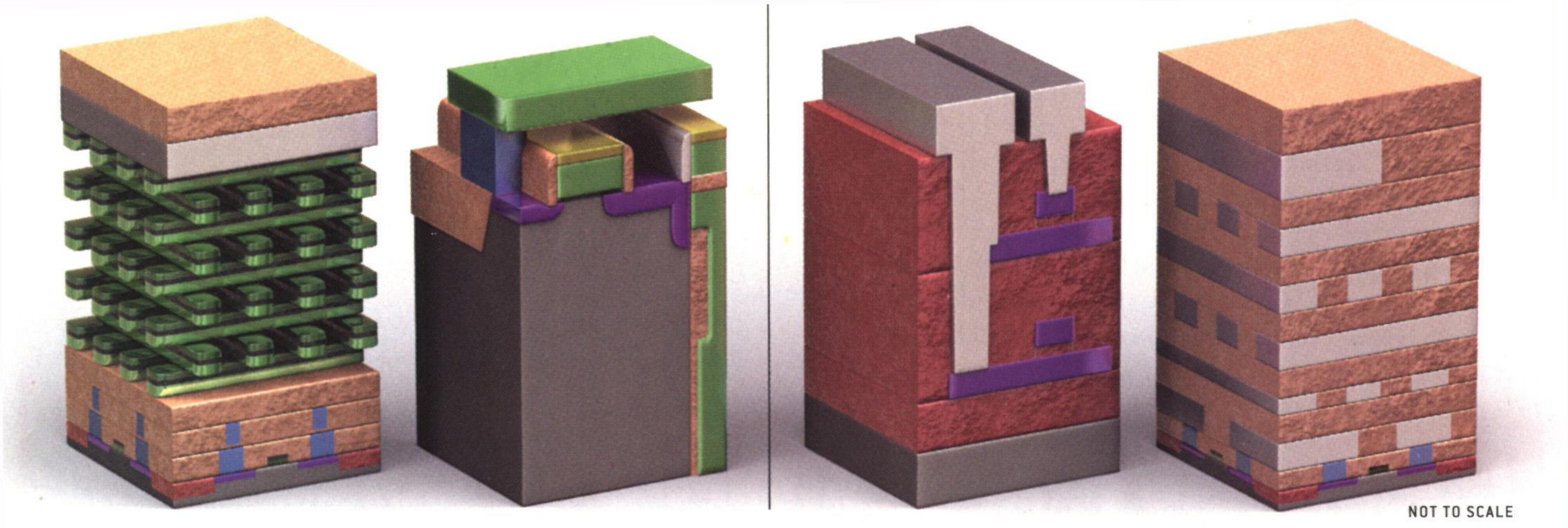
Single-phase, miniaturized convective cooling



SEM section of two-level jet plate. Water flow is indicated by blue arrows.

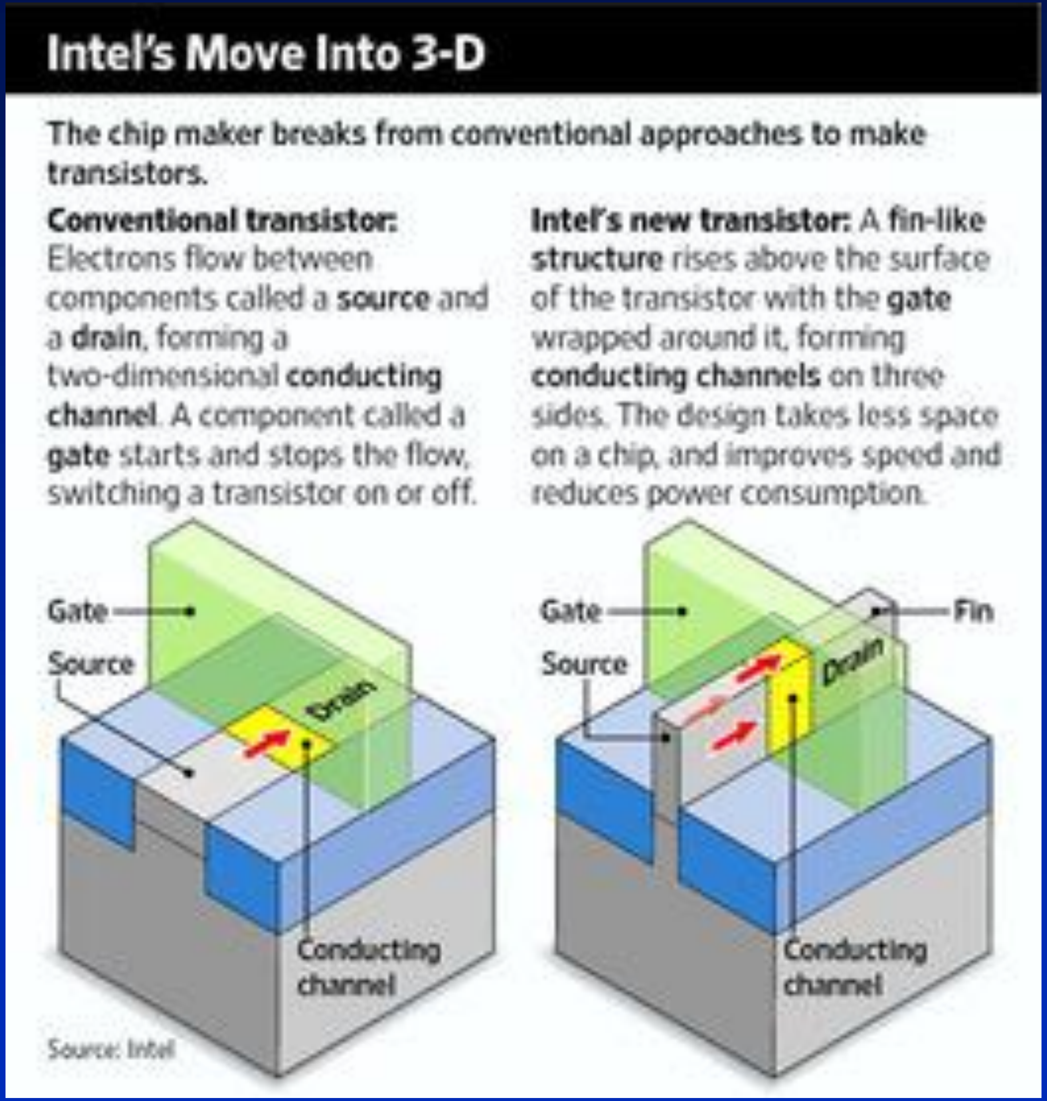
Interior Structure of 3-D chips

3-D Volatile Memory [Matrix Semiconductor]	2-D Random-Access Memory [IBM 256-Megabit]	3-D Logic Circuit [Lab Prototype]	2-D Microprocessor [Advanced Micro Devices Athlon]



⬛ Monosilicon substrate 🟫 Insulators ⬜ Aluminum wires 🟩 Polysilicon 🟦 Tungsten plugs 🟪 Ion-doped silicon 🟥 Isolation oxides 🟨 Silicide

Intel’s 3D Transistor 2011



Intel’s Move Into 3-D

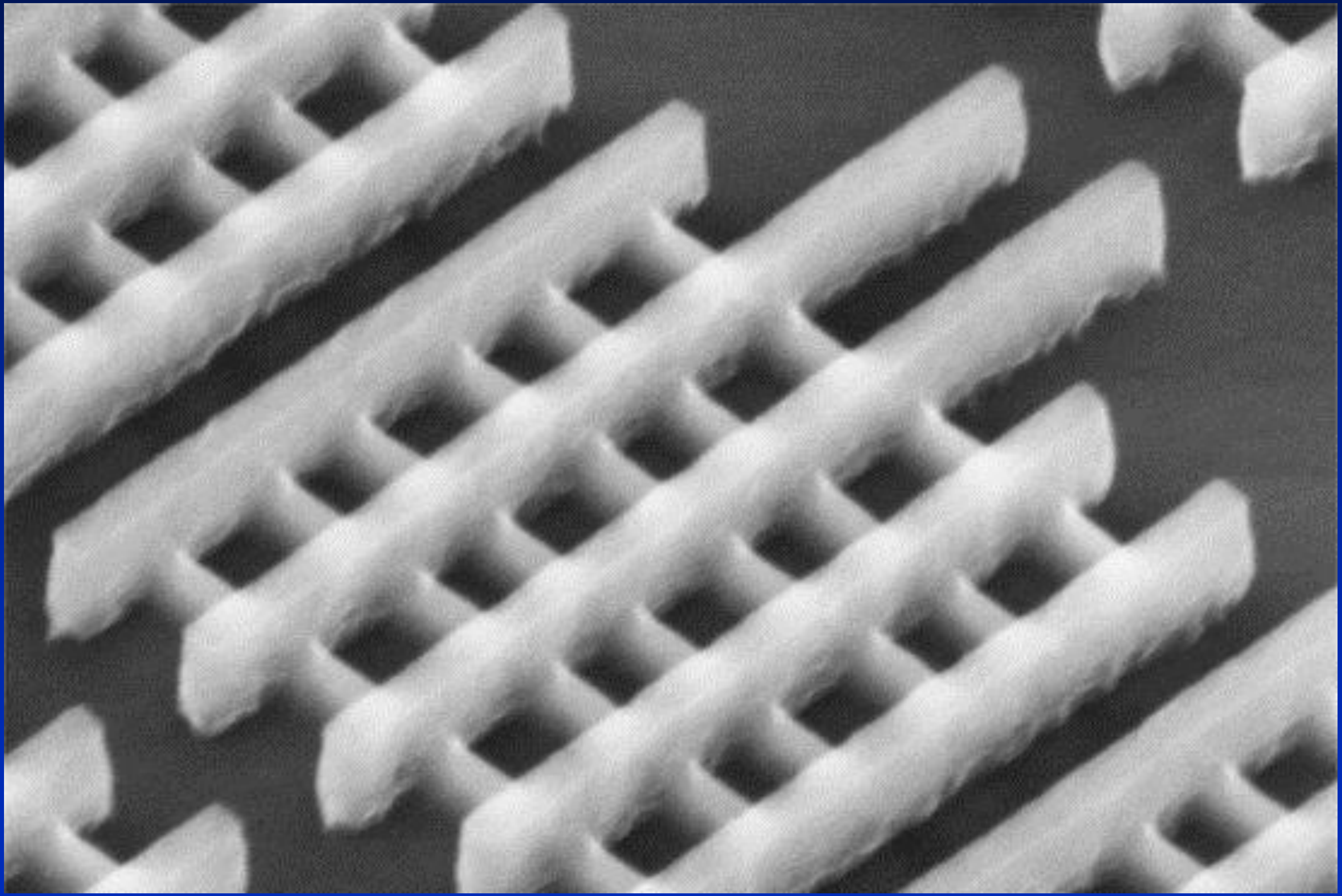
The chip maker breaks from conventional approaches to make transistors.

Conventional transistor: Electrons flow between components called a **source** and a **drain**, forming a two-dimensional **conducting channel**. A component called a **gate** starts and stops the flow, switching a transistor on or off.

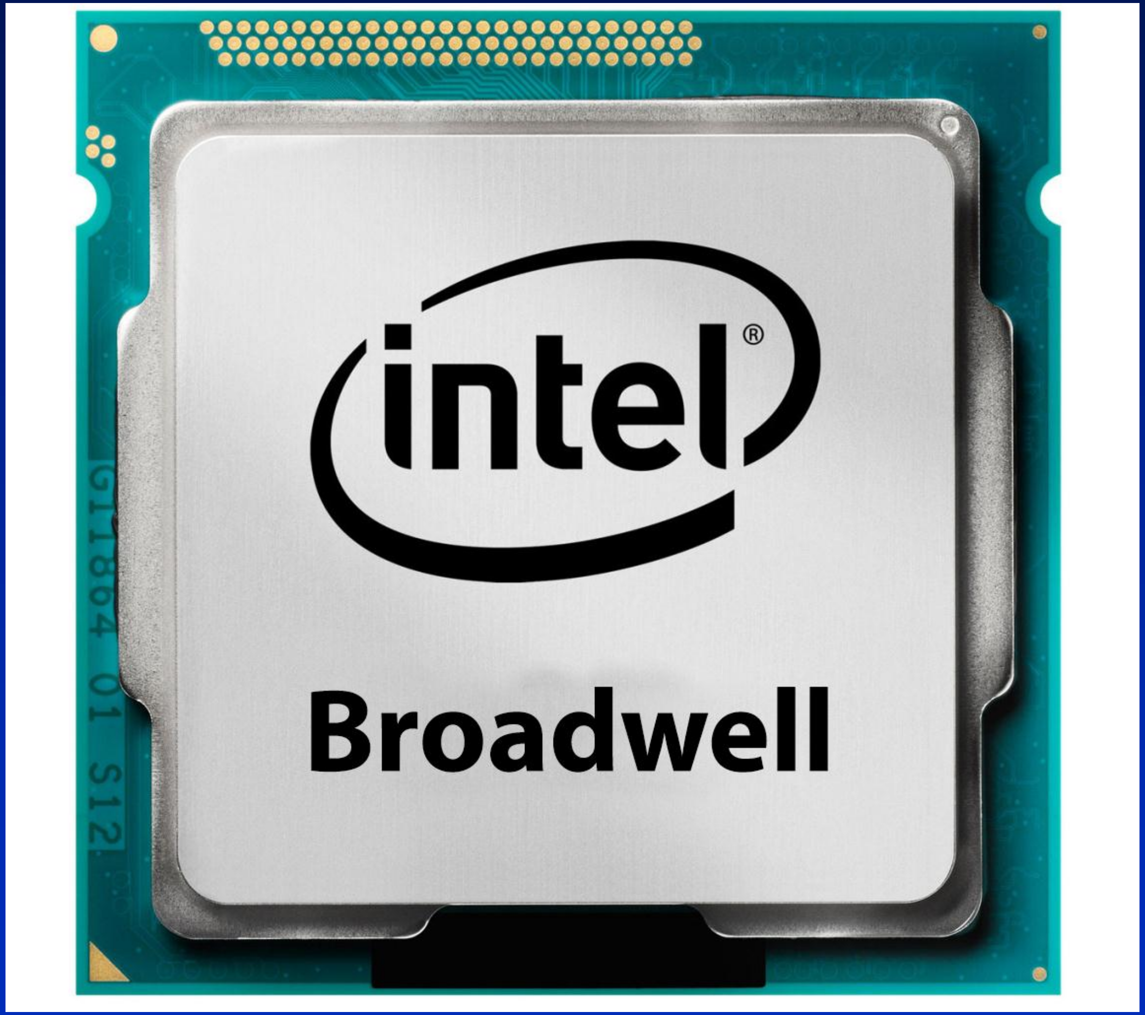
Intel’s new transistor: A **fin-like structure** rises above the surface of the transistor with the **gate** wrapped around it, forming **conducting channels** on three sides. The design takes less space on a chip, and improves speed and reduces power consumption.

Source: Intel

Intel’s 22nm 3D tri-gate transistor

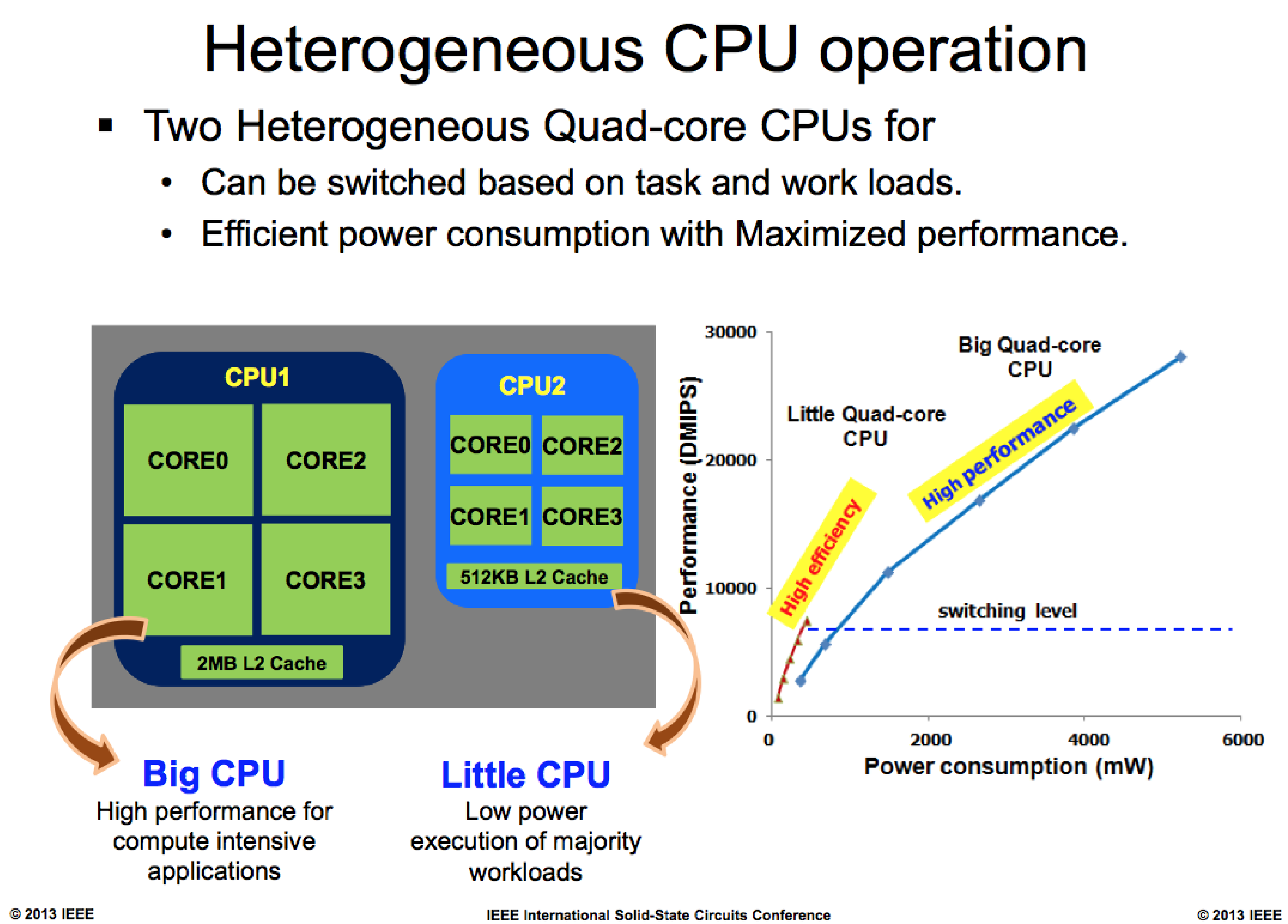


Intel 5th Generation i7 Chip (Broadwell) - 2014



- 4 core/8 threads for desktop and mobile solutions
- 14nm Hi-K+ process
- 3D tri-gate transistors
- Predicted 30% improvement in power consumption

Samsung Galaxy Exynos 5 Octa



Heterogeneous CPU operation

- Two Heterogeneous Quad-core CPUs for
  - Can be switched based on task and work loads.
  - Efficient power consumption with Maximized performance.

Big CPU
High performance for compute intensive applications

Little CPU
Low power execution of majority workloads

IEEE International Solid-State Circuits Conference

Samsung Galaxy S5 April 2014

- Samsung Exynos 5 Octa 5422, (8 cores)
- Heterogeneous ARM architecture (2.1 GHz Cortex-A15 quad, 1.5 GHz Cortex-A7 quad core)
- 2 GB DDR3 Ram
- 64 GB storage

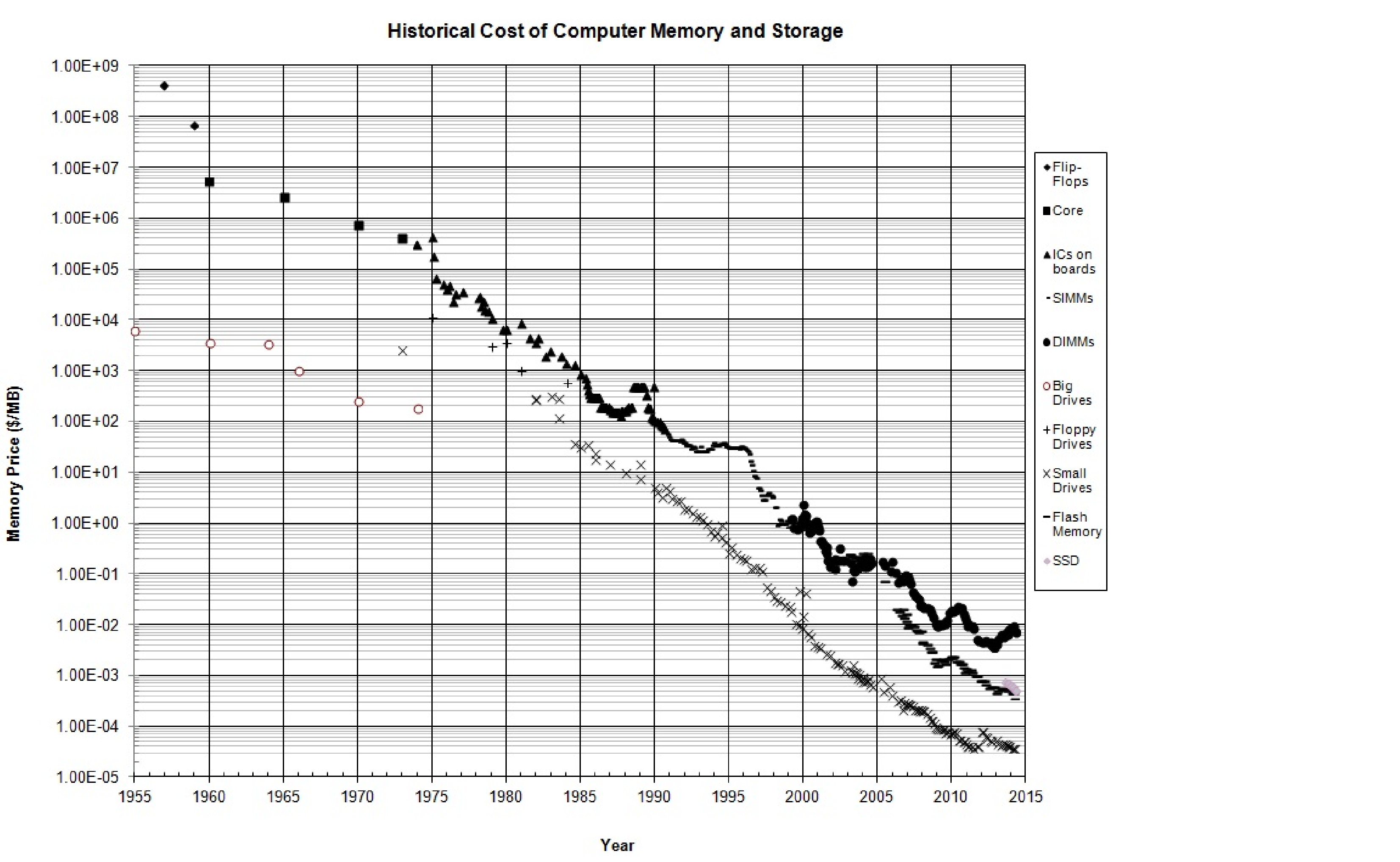
Samsung Galaxy S6 2016

- Screen Size: 5.1 inch screen
- Screen Resolution: 1440 x 2560 pixels (~577 ppi)
- Battery Life: 23 hour talk time
- Feature: 4G LTE
- Operating System: Android
- Camera Resolution: 16 MP
- Weight: 4.9 ounce
- Memory: 32 GB
- 64-bit Exynos 7420 Processor
- 14 nm FinFET technology

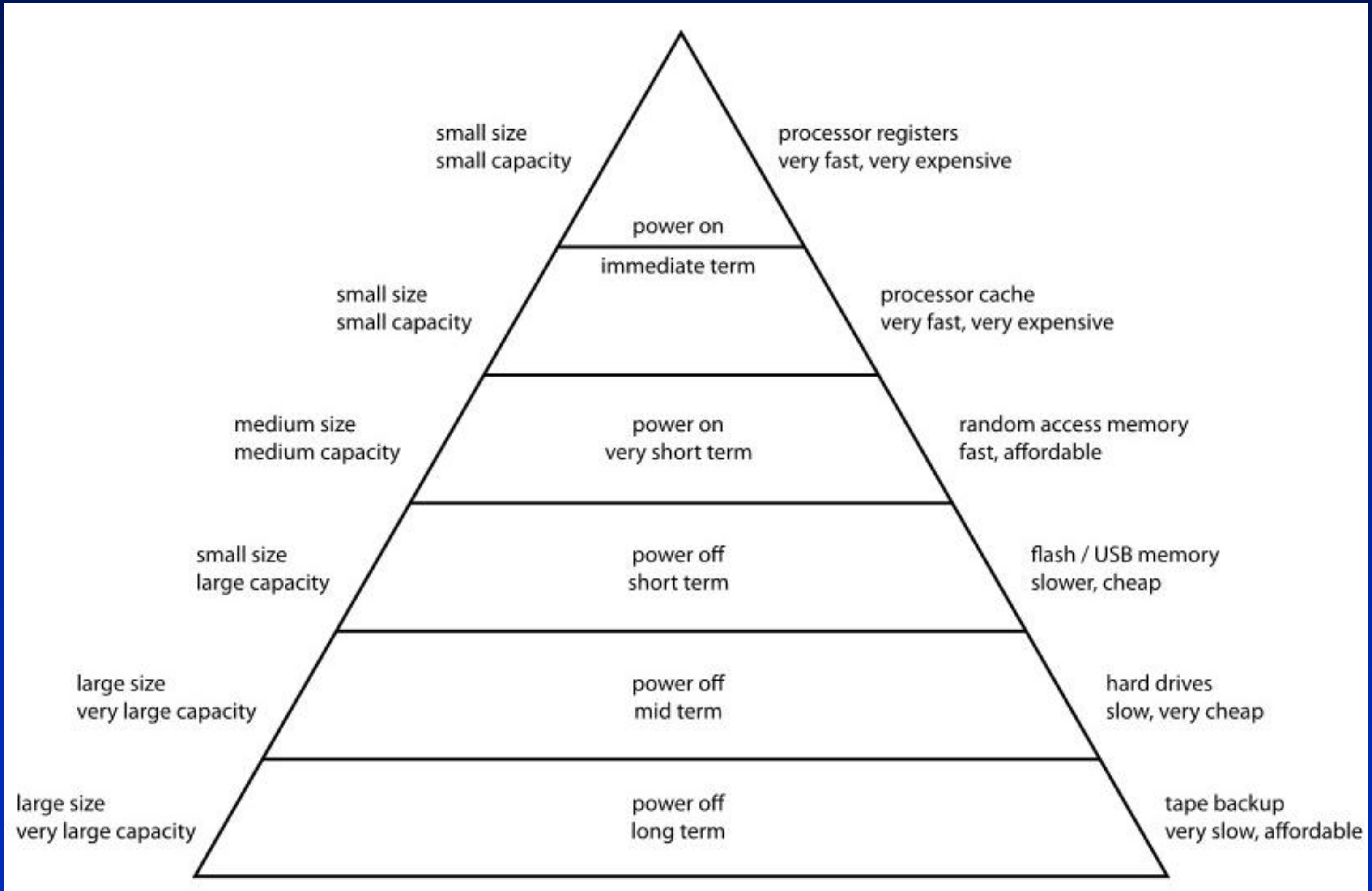
Intel's $7B Investment 2017

Wall Street Journal, Feb. 8th, 2017

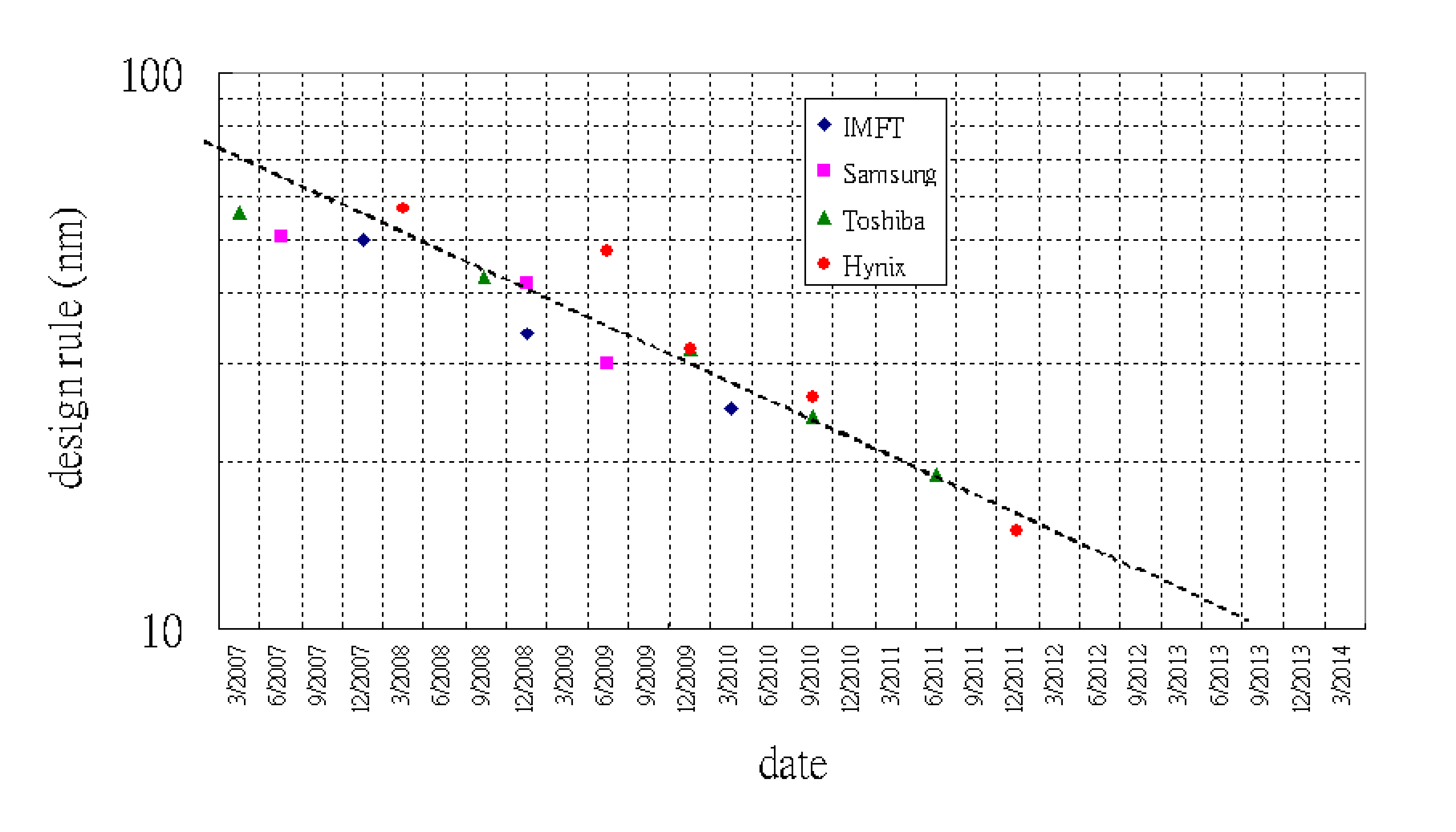
Cost of Computer Memory and Storage



Computer Memory Hierarchy



Flash Scalability

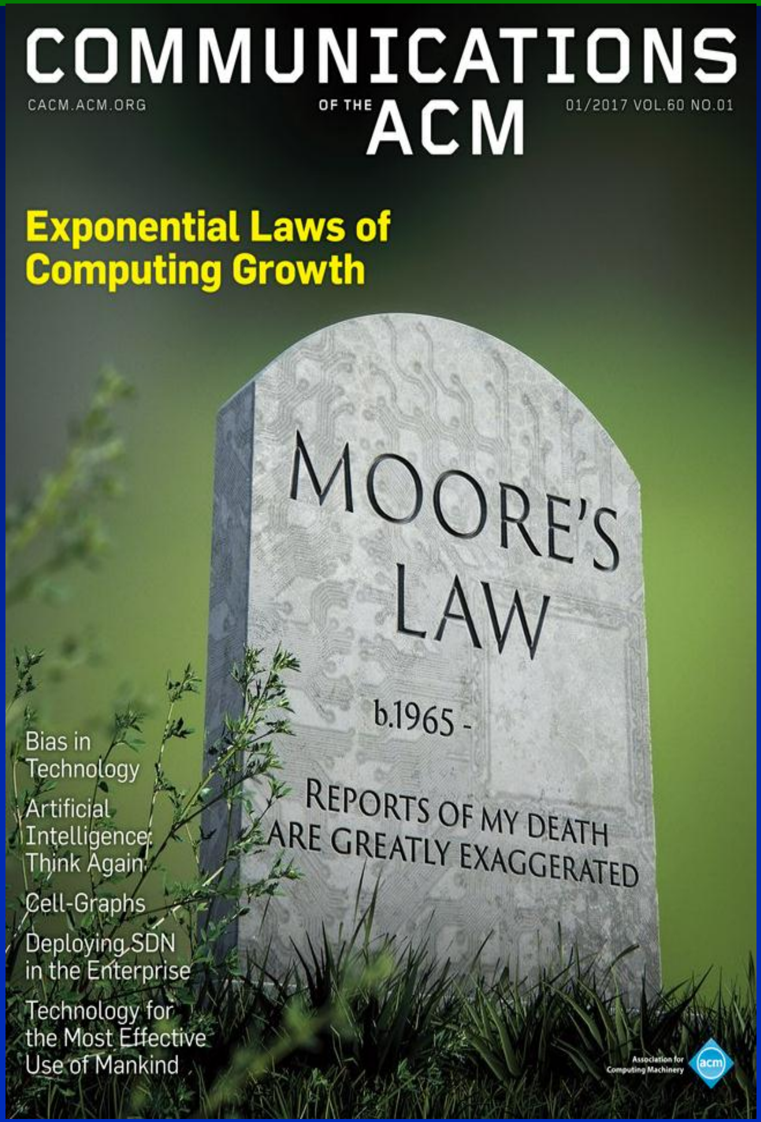


"Every economic era is based on a key abundance and a key scarcity."

George Gilder,
Forbes ASAP, 1992

What are the key scarcities?

Exponential Laws of Computing Growth



COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

01/2017 VOL.60 NO.01

Exponential Laws of Computing Growth

MOORE'S LAW
b.1965 -
REPORTS OF MY DEATH ARE GREATLY EXAGGERATED
Association for Computing Machinery
acm

Bias in Technology

Artificial Intelligence: Think Again

Cell-Graphs

Deploying SDN in the Enterprise

Technology for the Most Effective Use of Mankind

Dennings and Lewis, Jan. 2016

End