Purdue University Department of Physics: February 22, 2024

Essential Physics of the Ultimate MOSFET

and

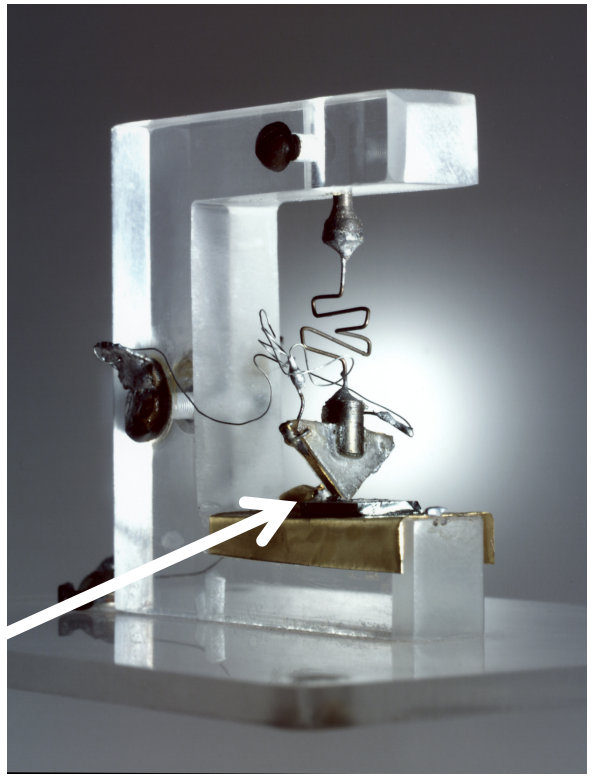
the next 20 years of semiconductor technology

Mark Lundstrom
School of ECE
Purdue University
lundstro@purdue.edu

PURDUE UNIVERSITY

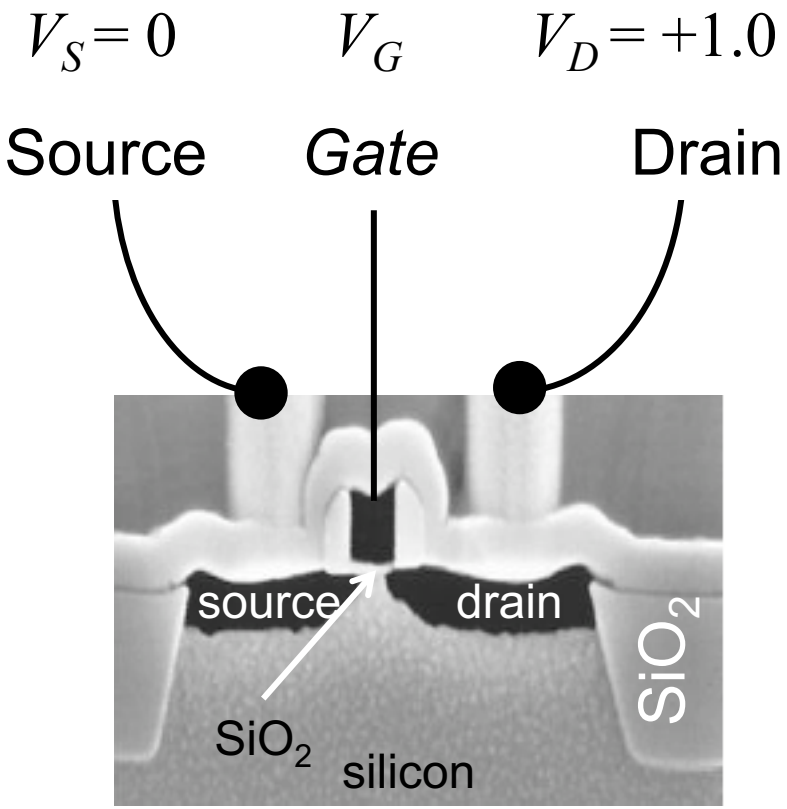
1947: The first transistor

Brattain, Bardeen, and Shockley, Bell Labs.



"The transistor was probably the most important invention of the 20th Century."
http://www.pbs.org/transistor/

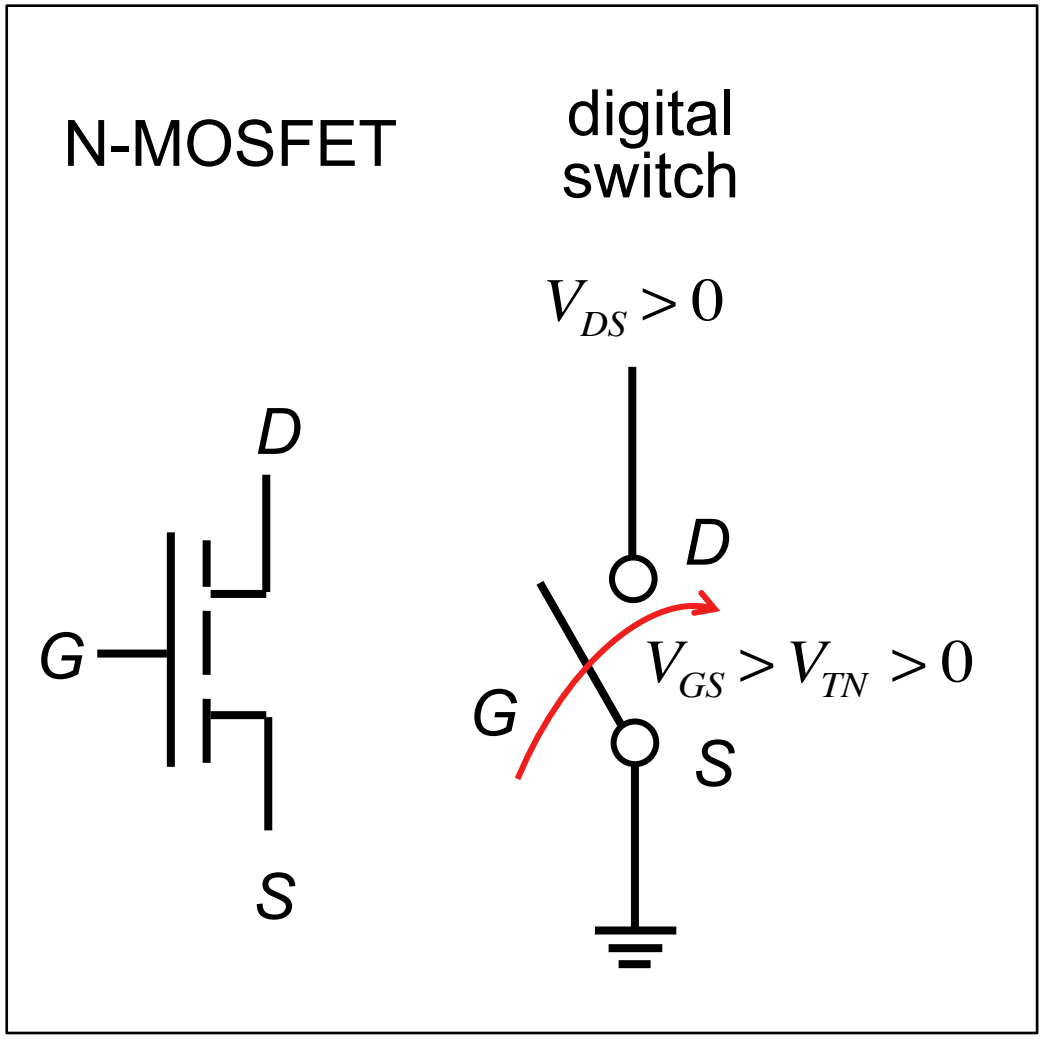
1960: The first Si MOSFET

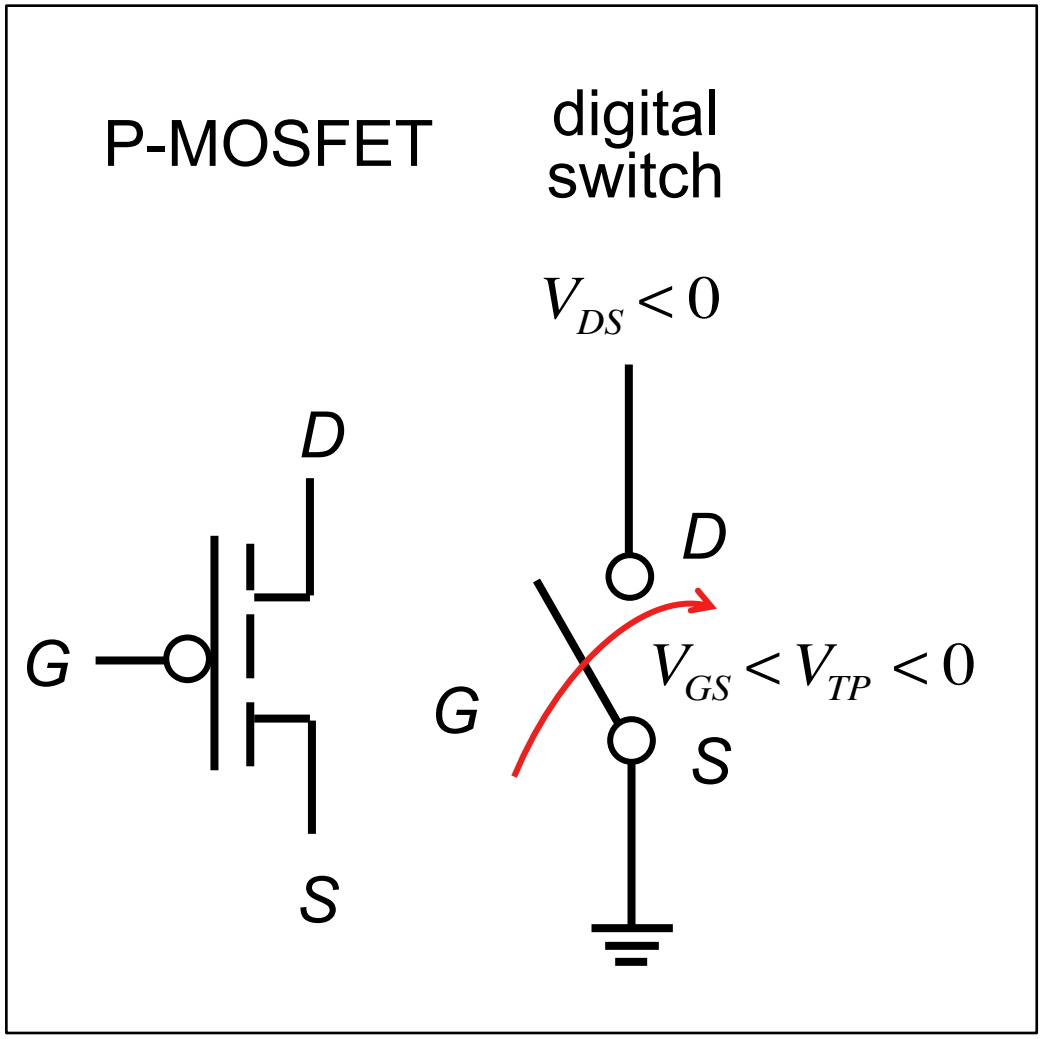


Bell Labs, 1959

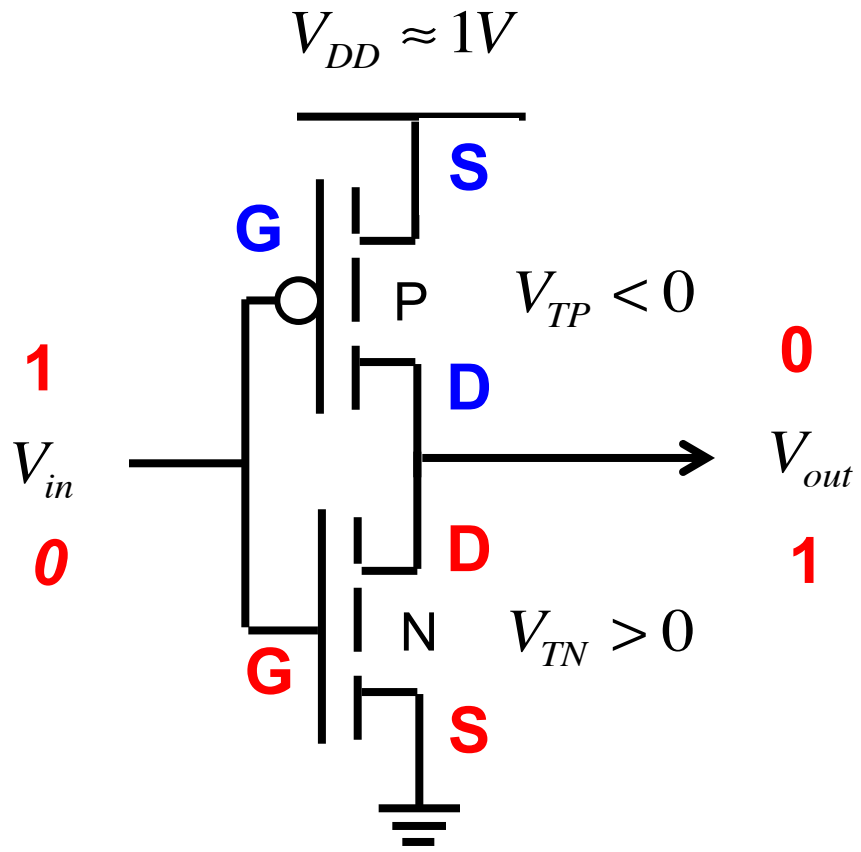
John M. Atalla
PhD Purdue University, 1949

Complementary MOSFETs



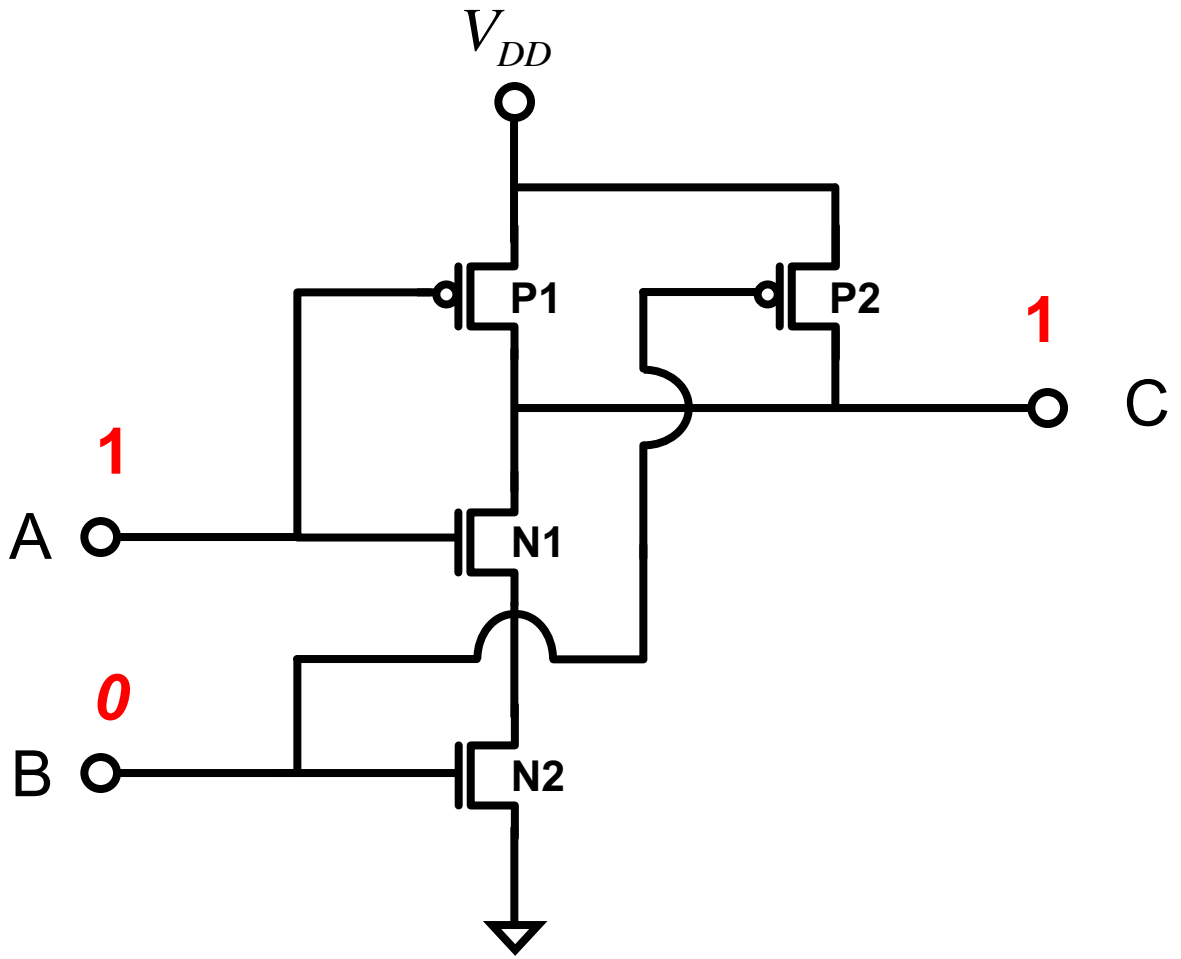


CMOS inverter



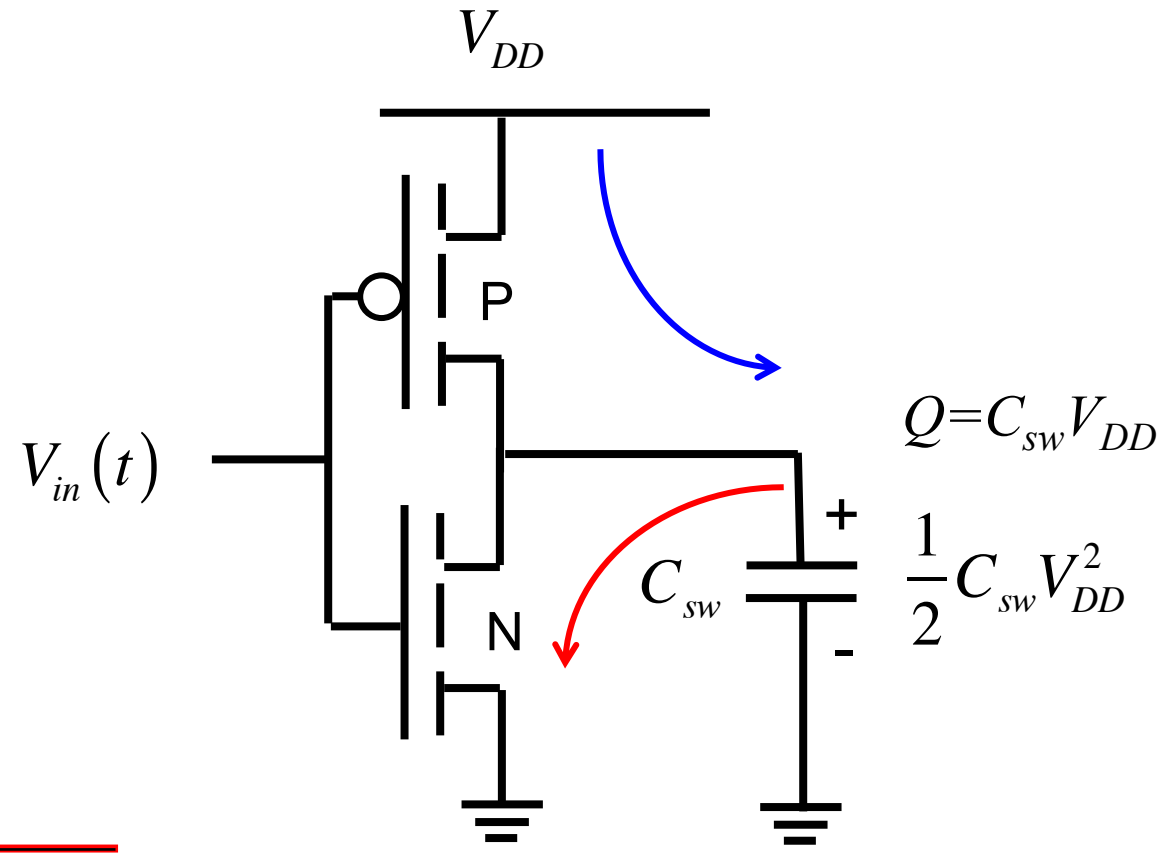
CMOS logic: 2-input NAND gate

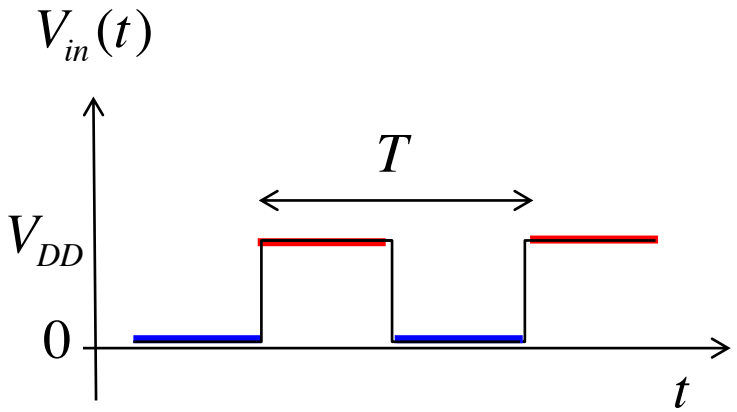
AND		
A	B	C
0	0	0
0	1	0
1	0	0
1	1	1



NAND		
A	B	C
0	0	1
0	1	1
1	0	1
1	1	0

CMOS speed and power





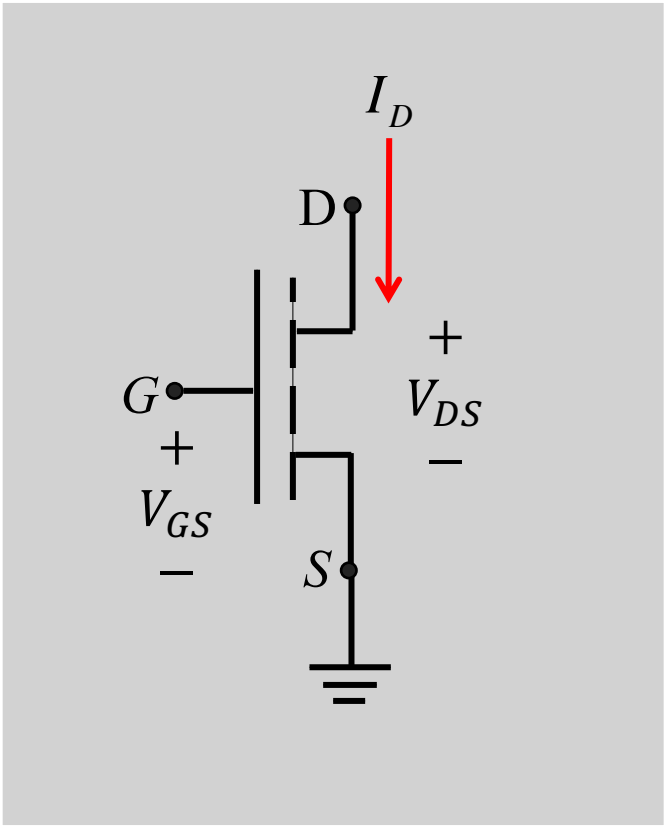
$$\tau = \frac{C_{sw}V_{DD}}{I_{ON}}$$

$$P_{dynamic} = \alpha f C_{sw} V_{DD}^2$$

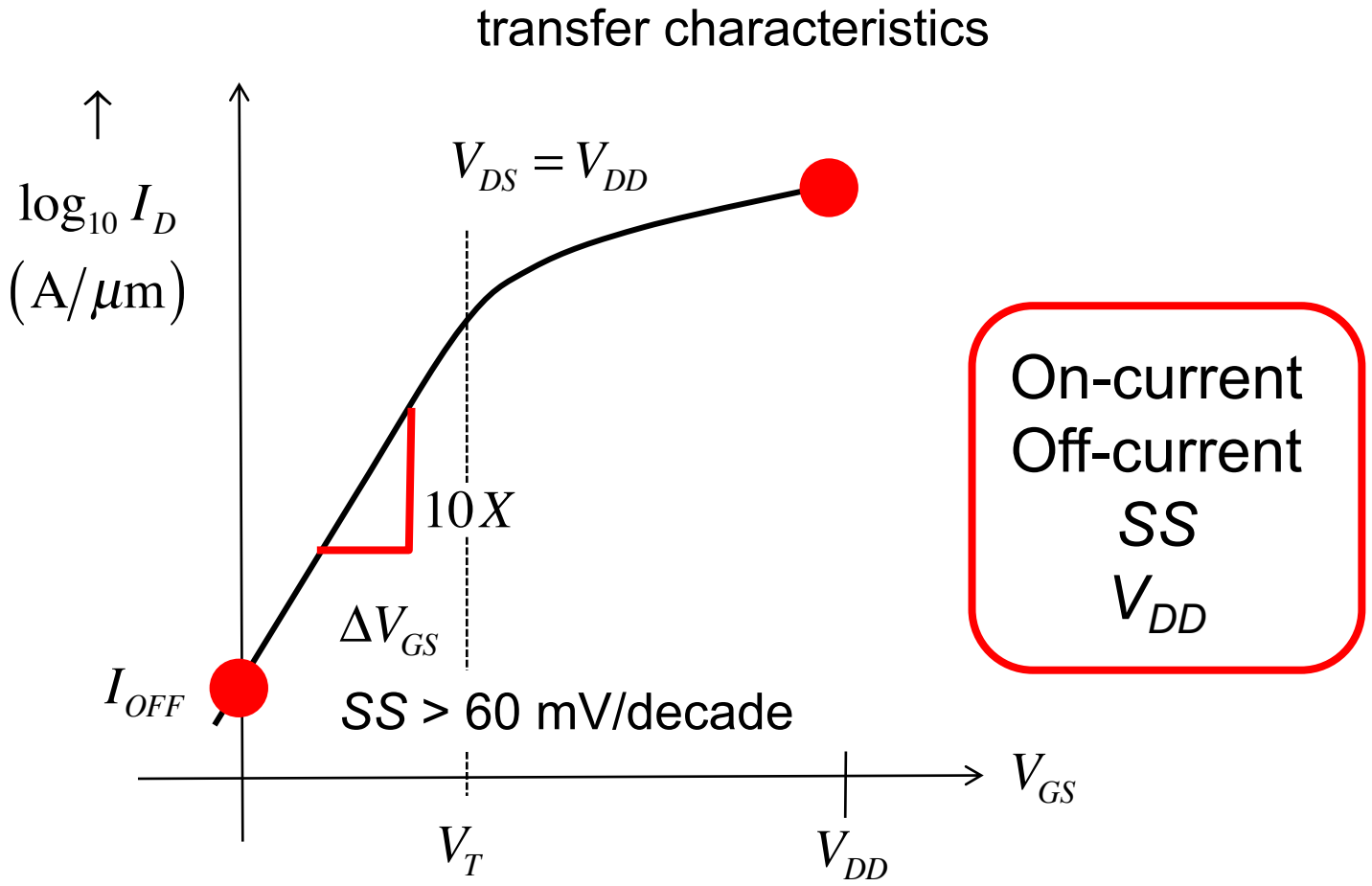
$$P_{static} = N_G I_{off} V_{DD}$$

On-current
Off-current
Power supply voltage

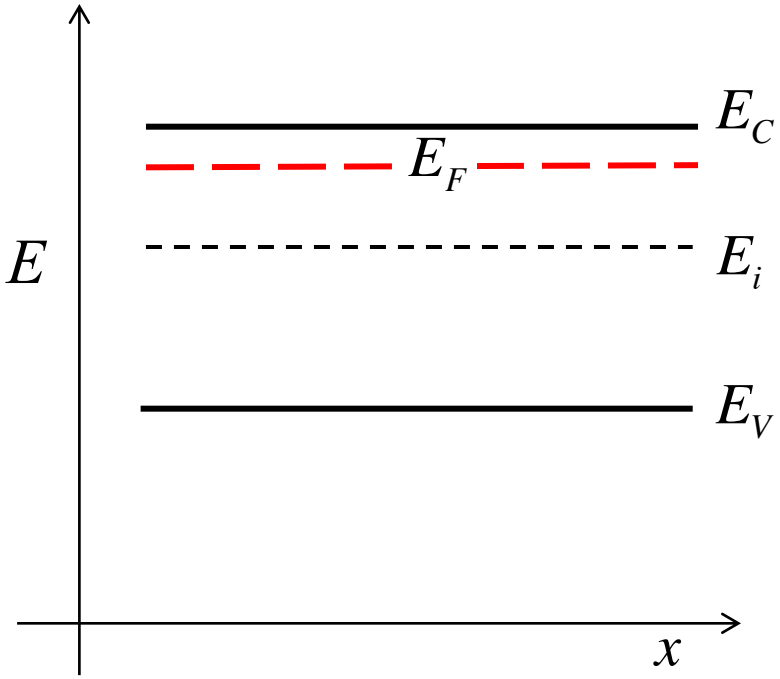
N-channel MOSFET

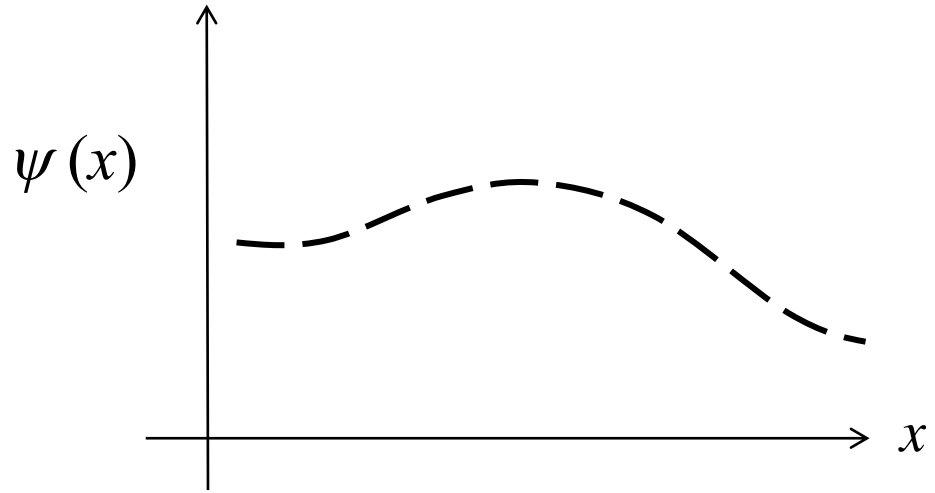


N-channel NMOS
(P-channel PMOS)



Energy band diagrams

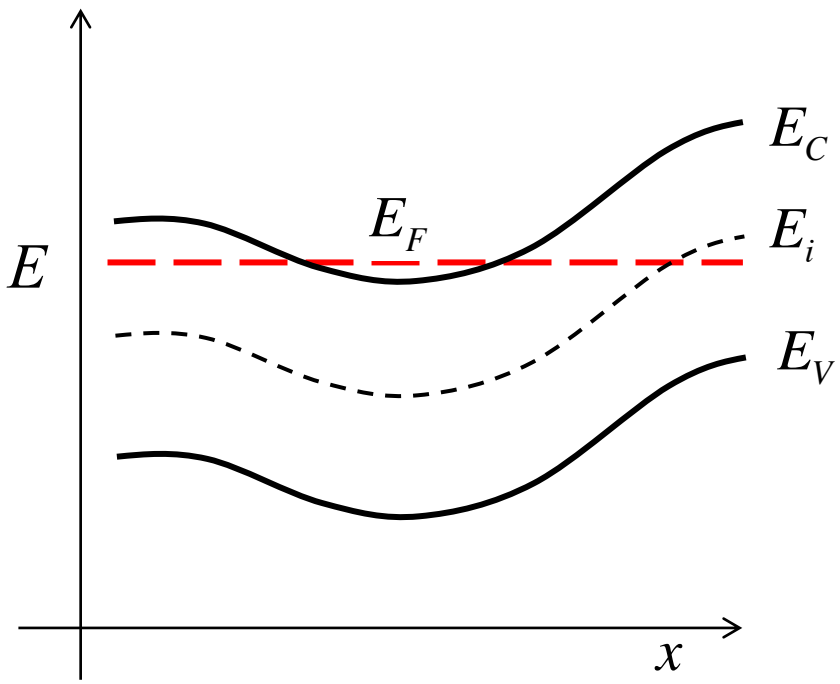


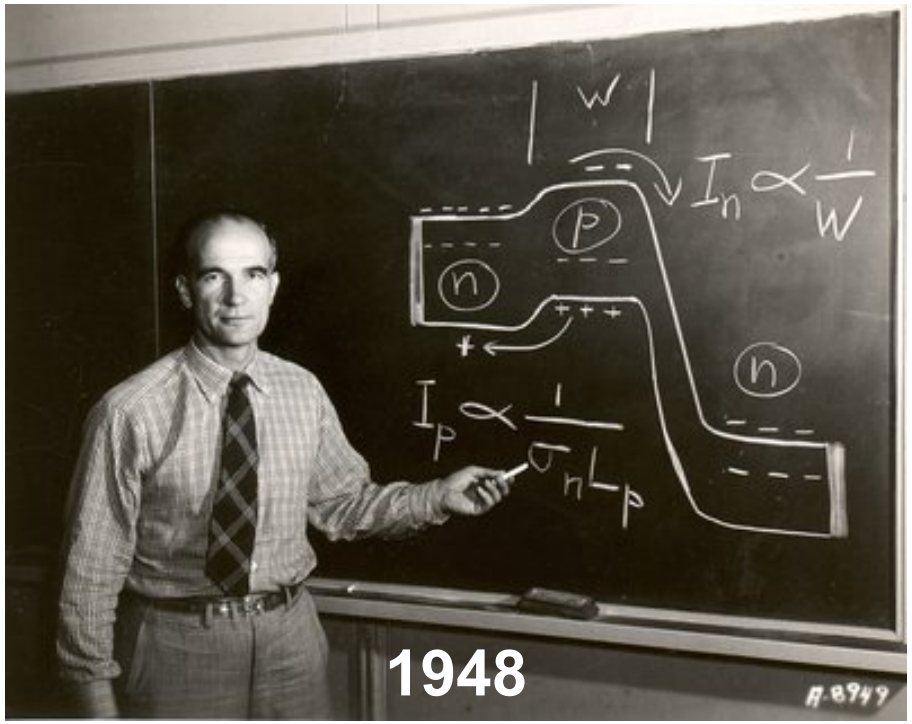


$$f_0(E) = \frac{1}{1 + e^{(E - E_F)/k_B T}}$$

$$E = -e\psi(x)$$

Energy band diagrams

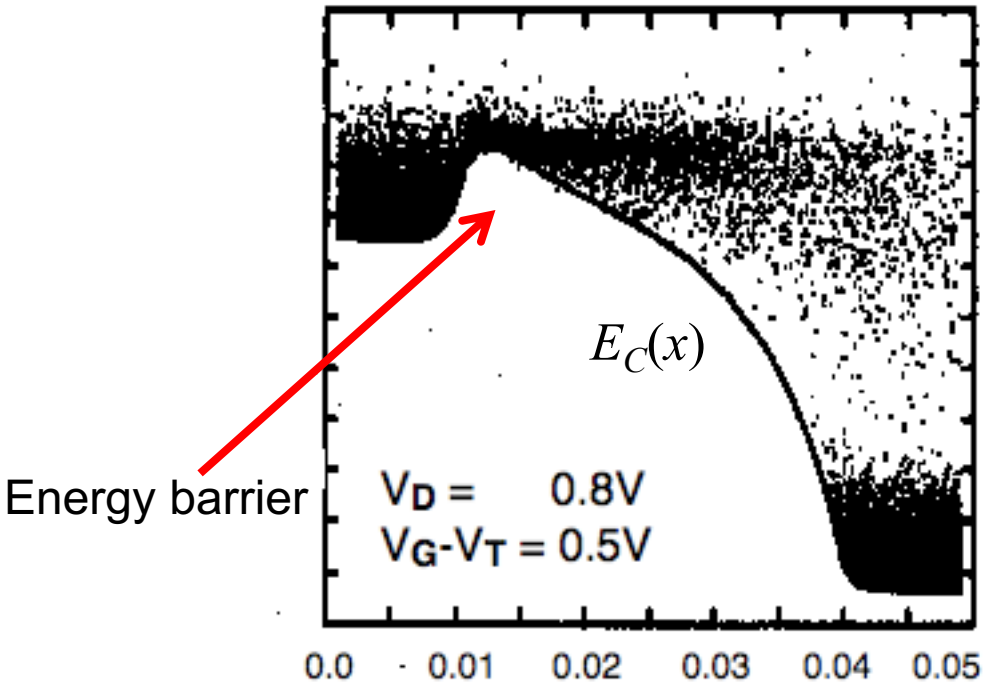




https://www.pbs.org/wgbh/americanexperience/features/silicon-timeline-silicon/

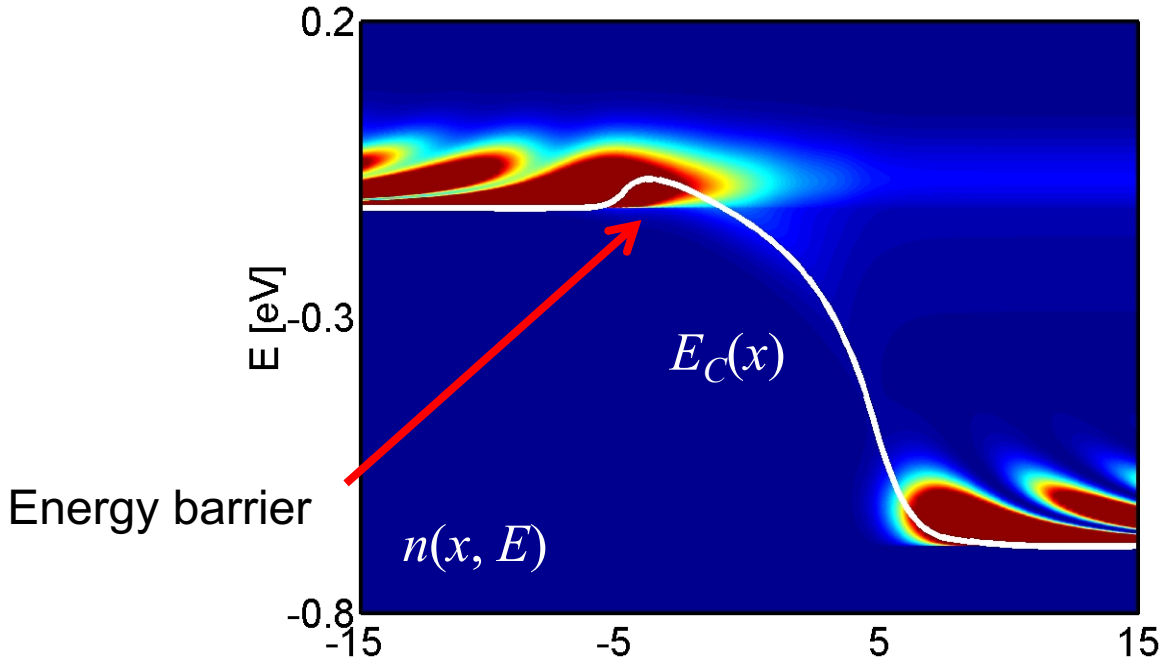
Electron transport in MOSFETs: Micro- to nano-scale

L = 30 nm



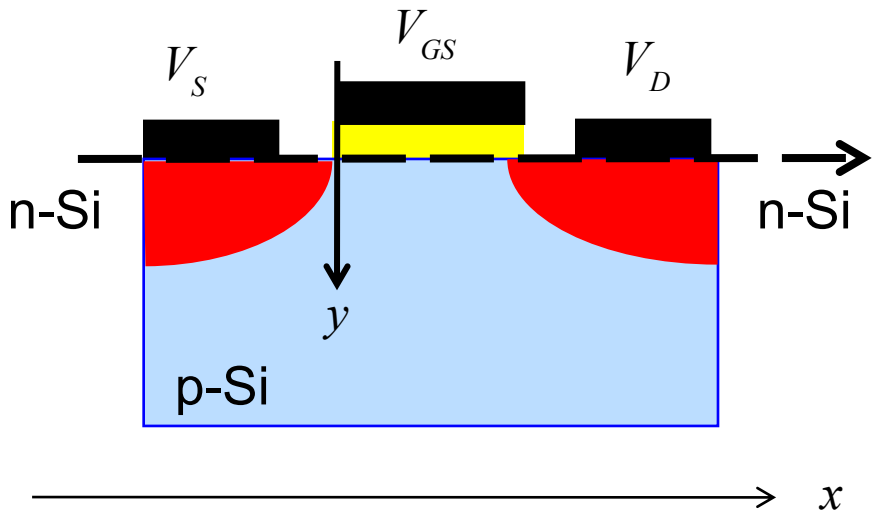
D. Frank, S. Laux, and M. Fischetti, Int. Electron Dev. Mtg., Dec., 1992.

L = 10 nm



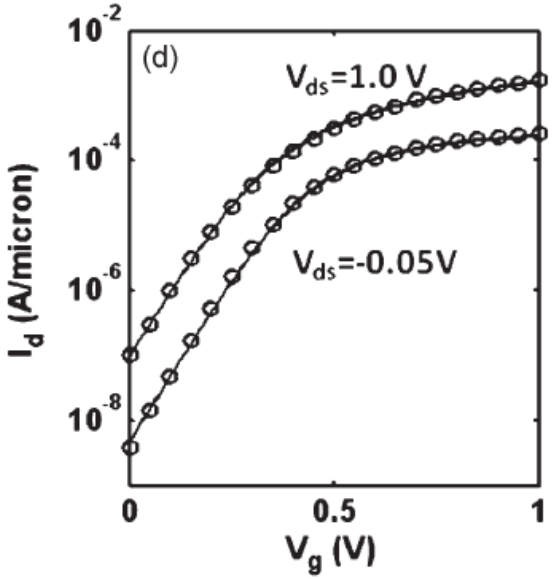
Ren, Venugopal, Goasguen, Datta, and Lundstrom, TED, **50**, p. 2185, 2003.

Energy band treatment of the MOSFET

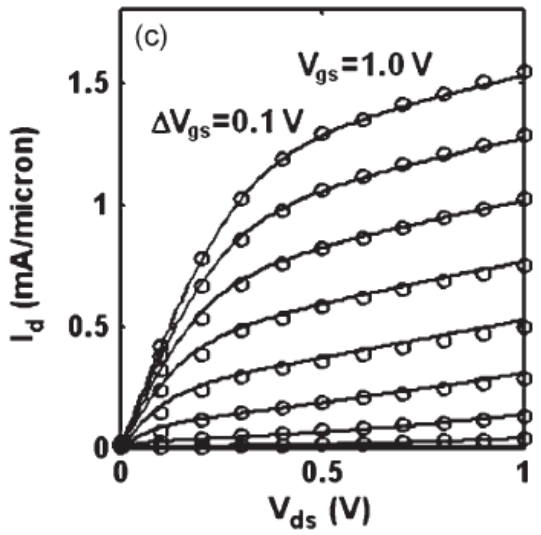


$$E_C(x) = E_{C0} - q\psi(x)$$

Transfer characteristics



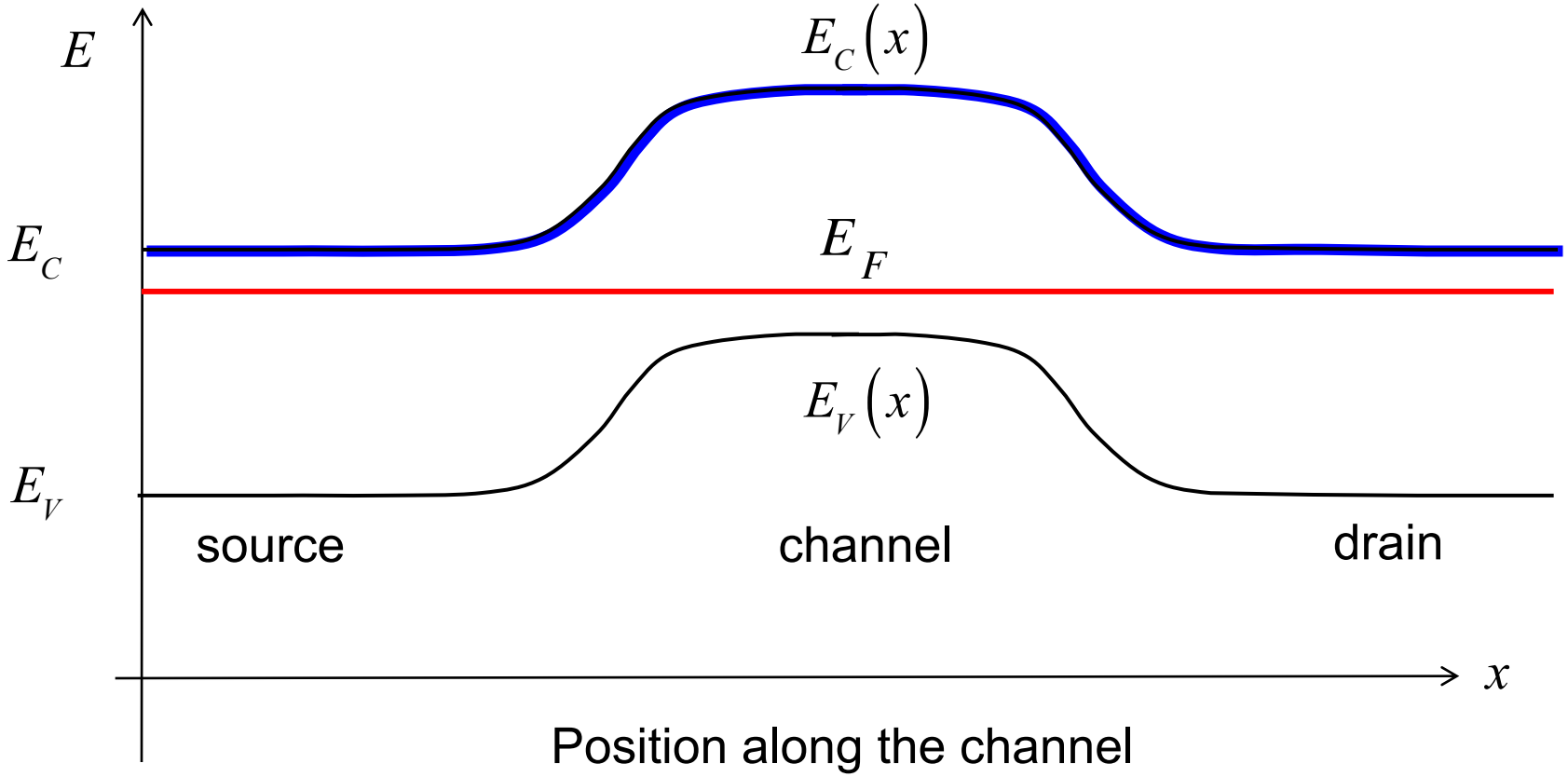
Output characteristics



32 nm technology

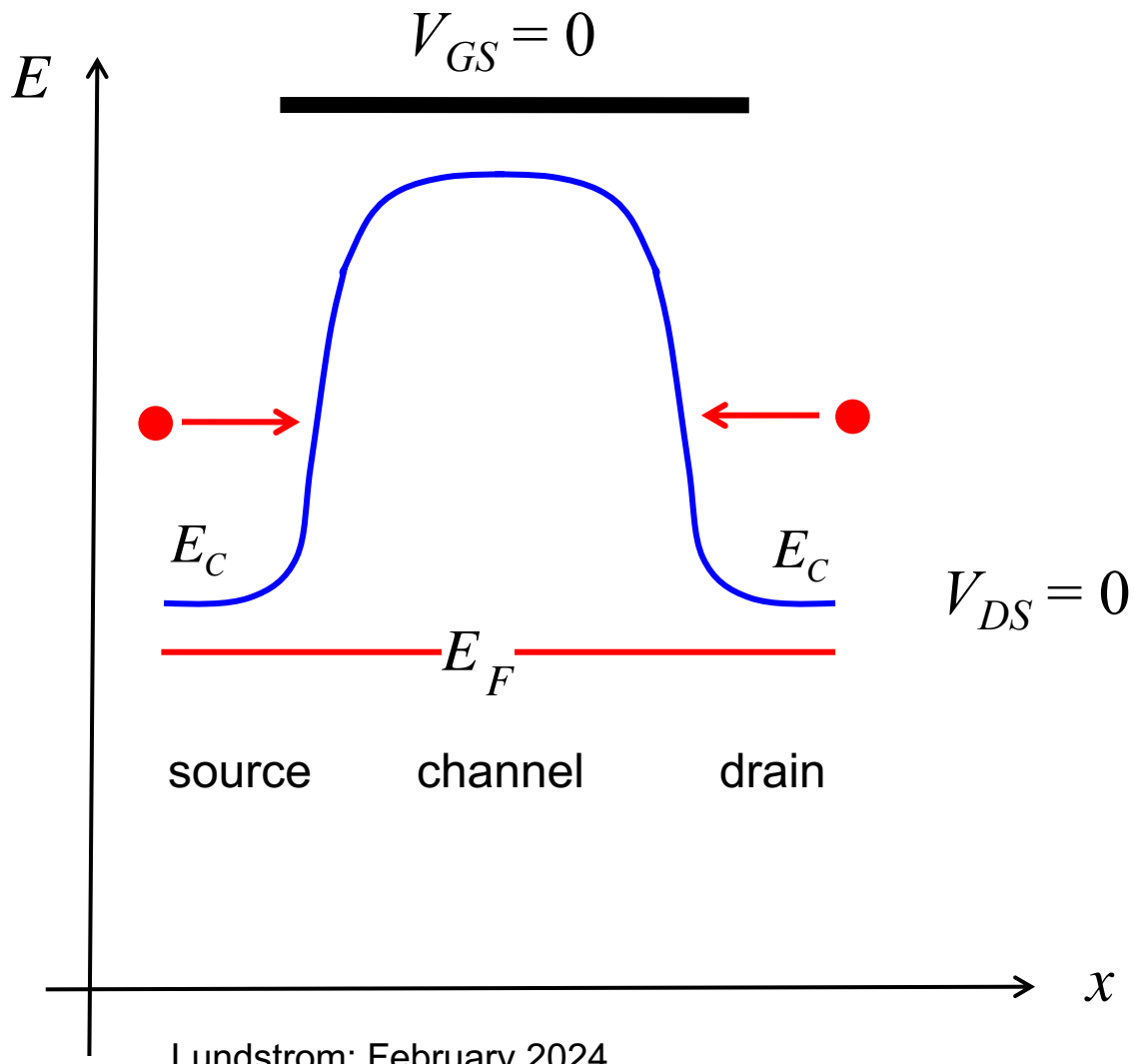
Equilibrium energy band diagram

Now, what effect does a gate voltage have?

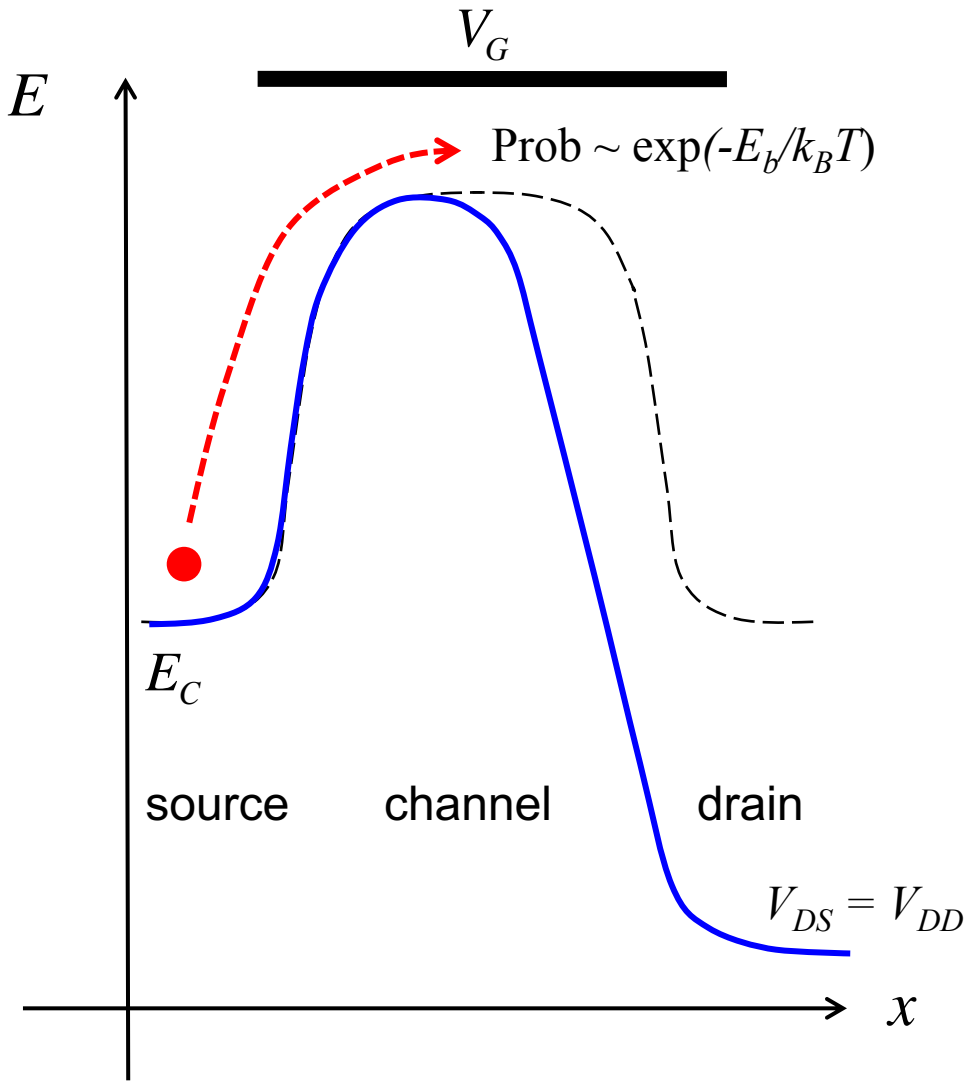


$V_{GS} = V_{DS} = 0$

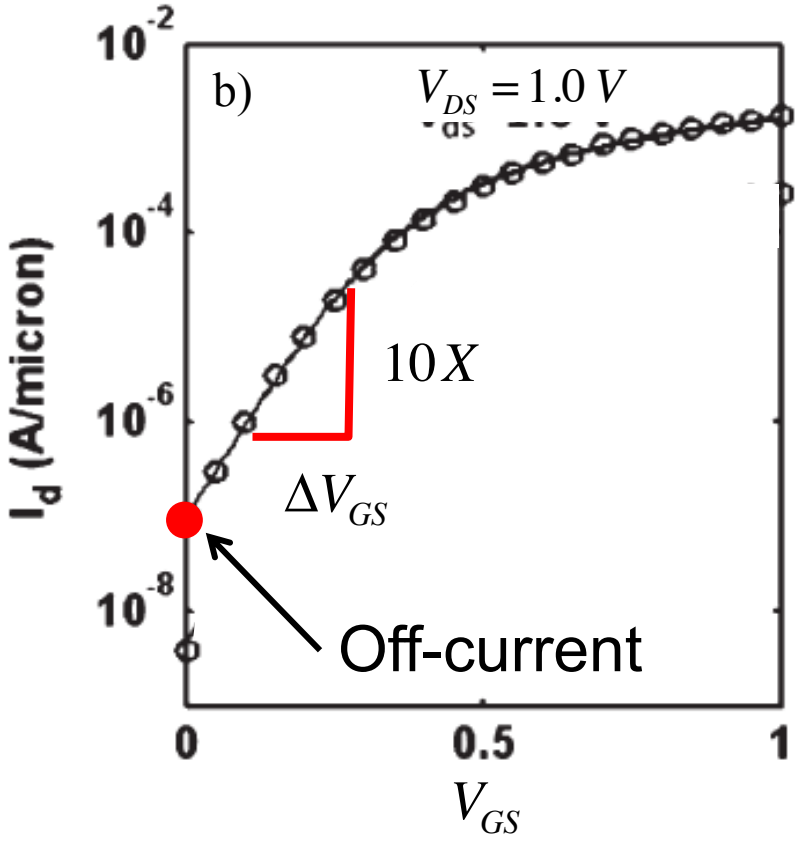
$$E_C(x) = E_{C0} - q\psi(x)$$



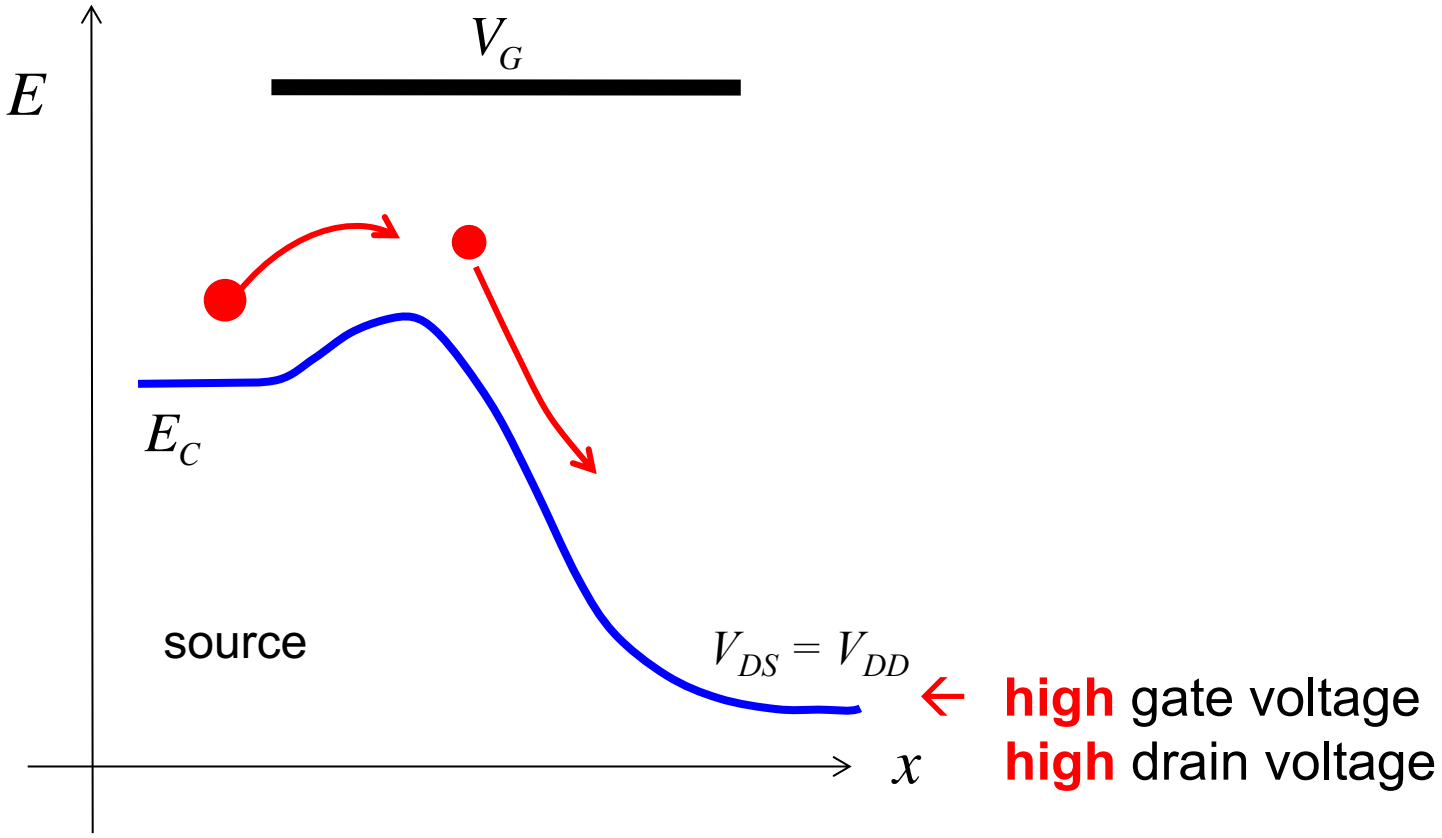
Off-state: $V_{GS} = 0$, $V_{DS} = V_{DD}$



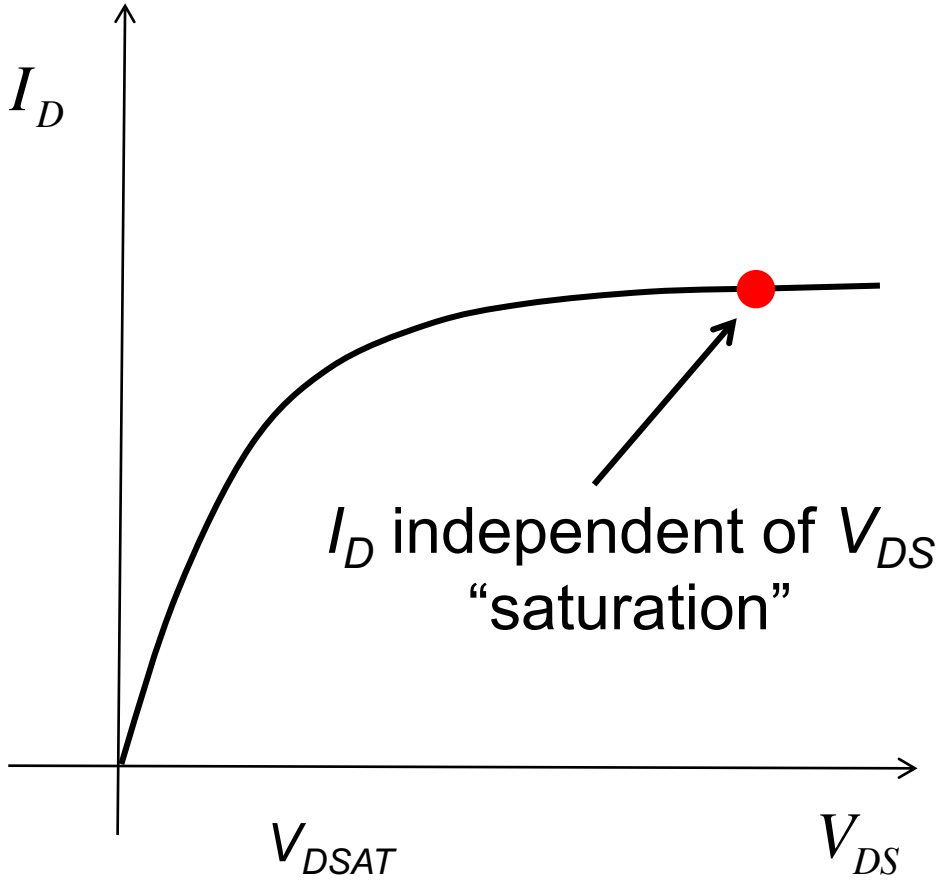
I_D vs. V_{GS} for a fixed V_{DS}



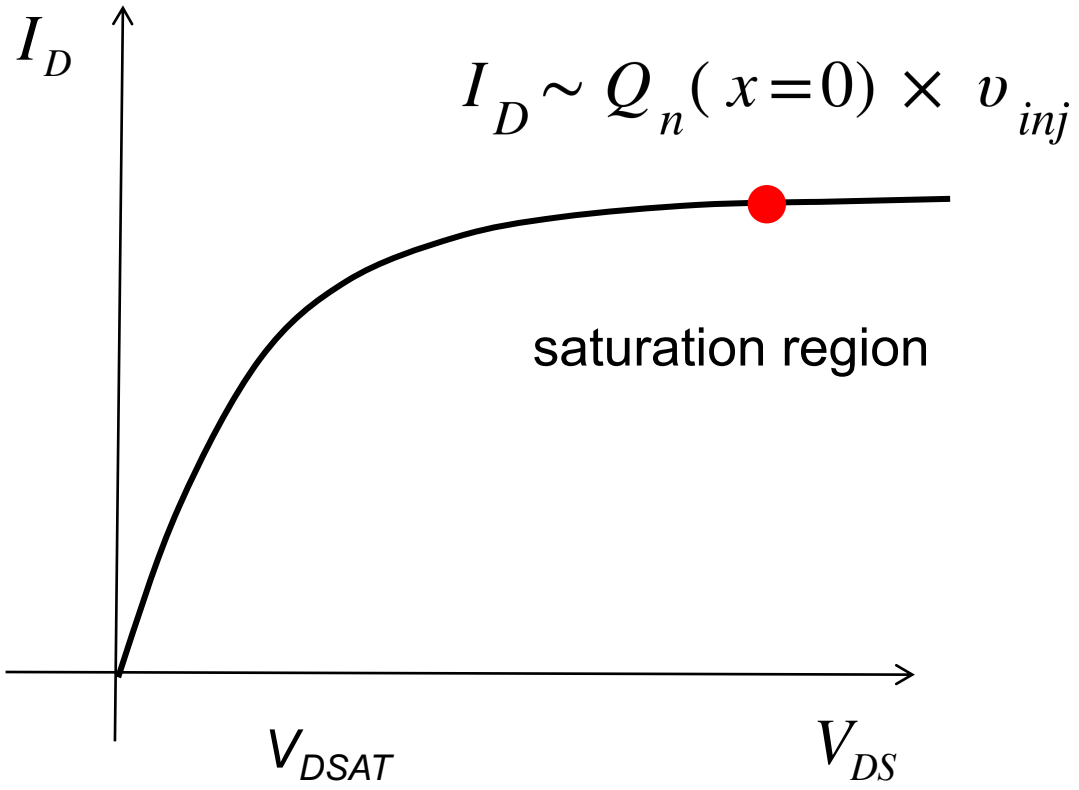
On-state: $V_{GS} = V_{DS} = V_{DD}$



← **high** gate voltage
high drain voltage



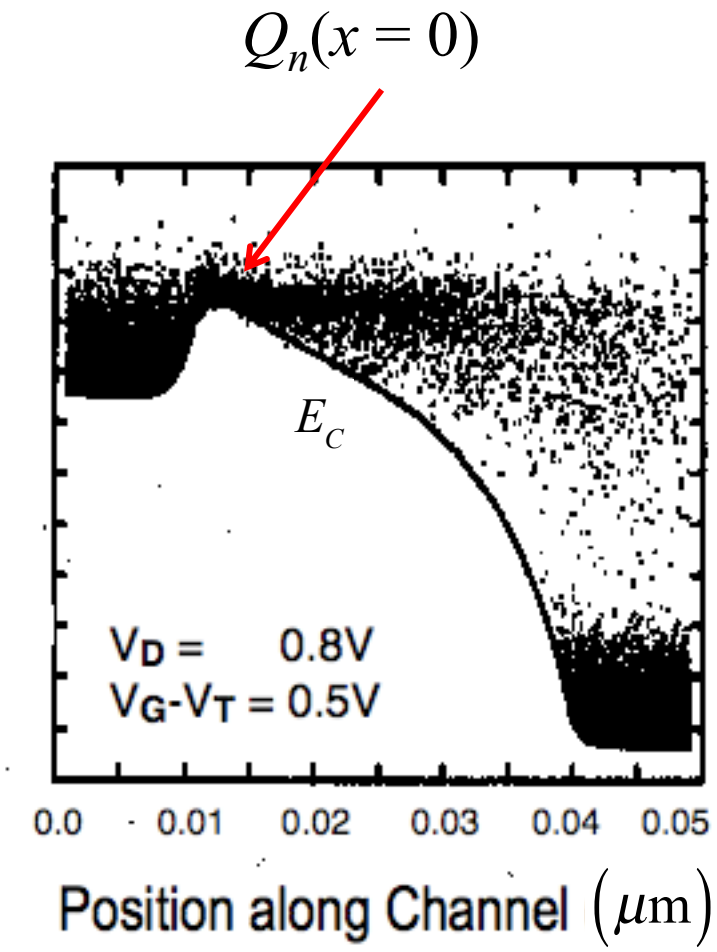
$V_{DS} > V_{DSAT}$ ("saturation" region)



flow is limited by the source

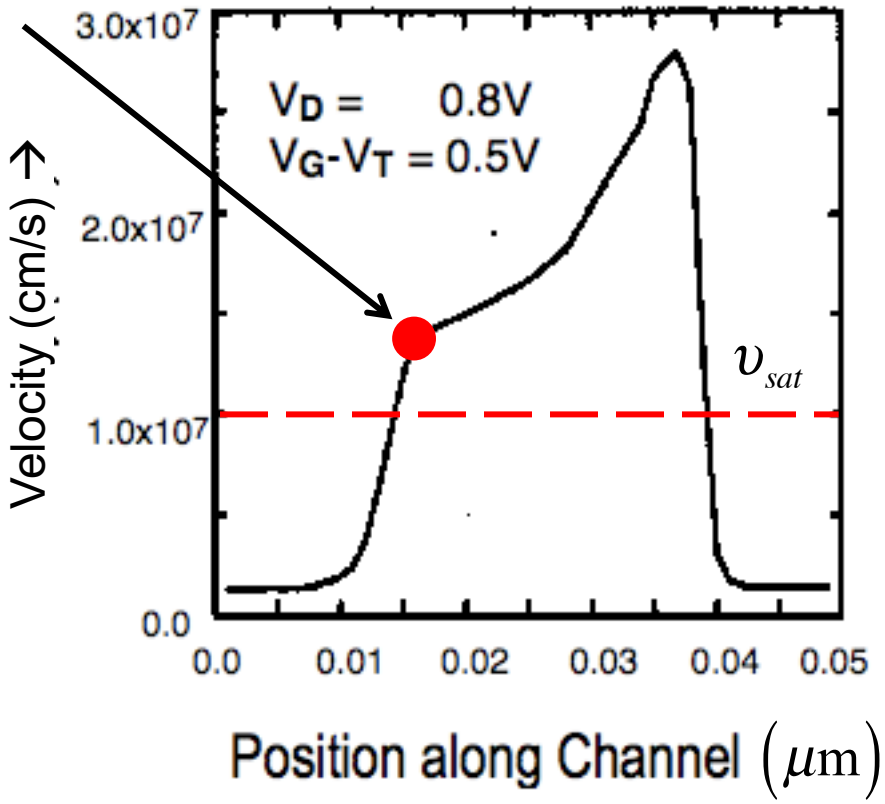
flow = quantity X velocity

High drain bias



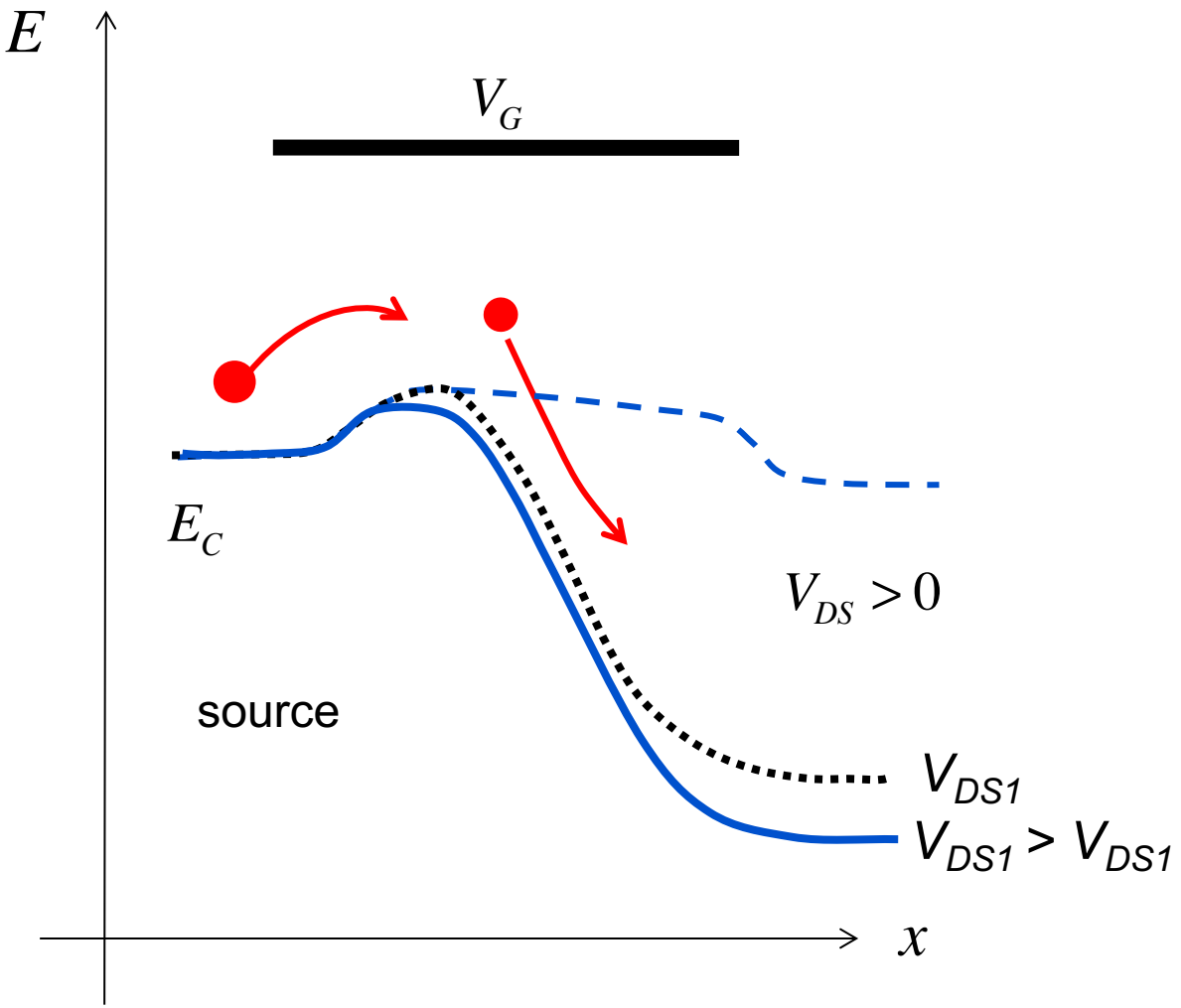
"injection velocity"

v_{inj}



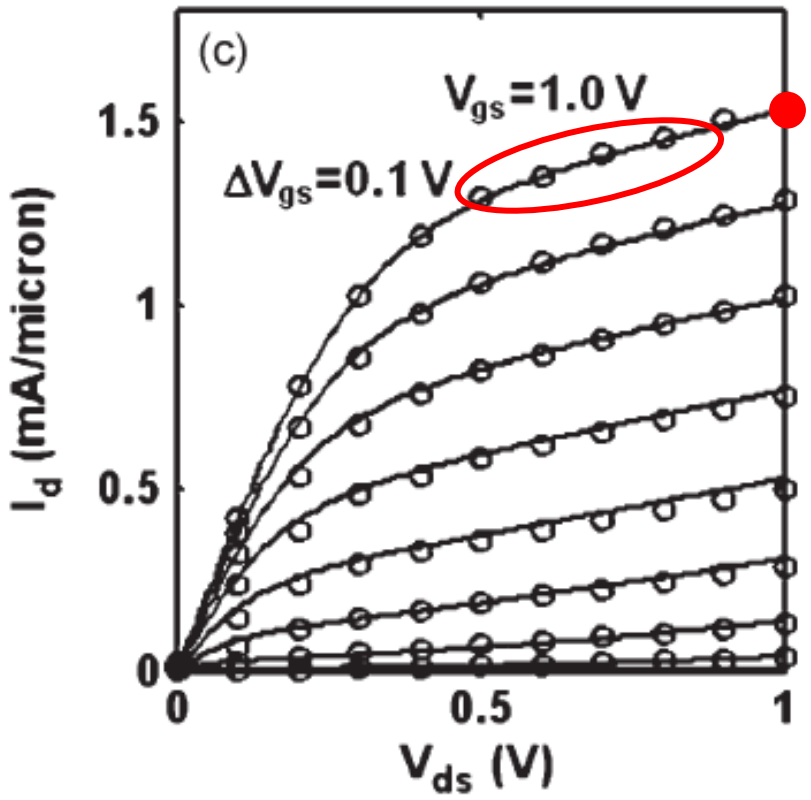
D. Frank, S. Laux, and M. Fischetti, Int. Electron Dev. Mtg., Dec., 1992.

Output conductance



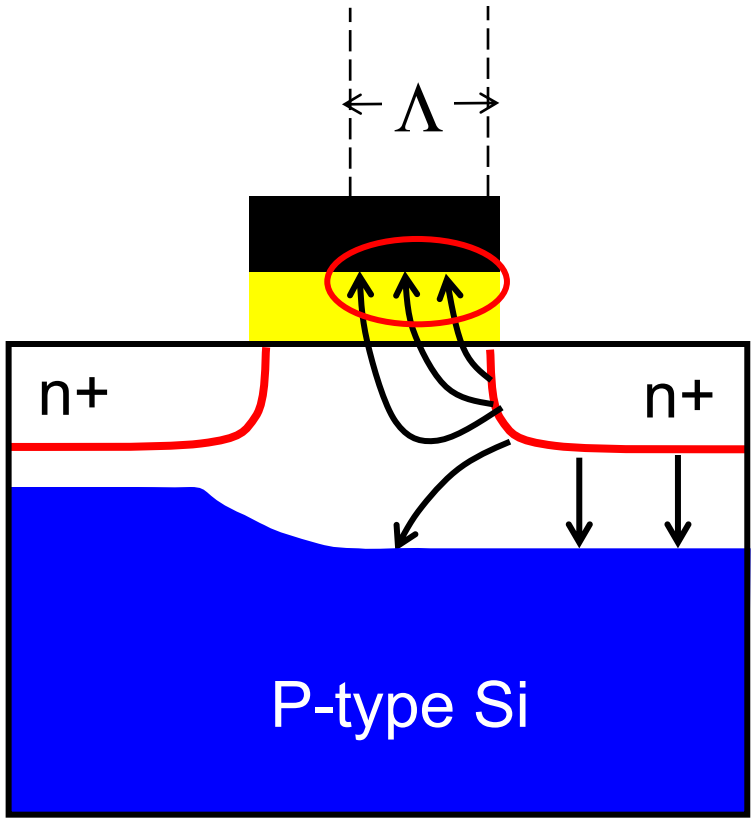
On-current

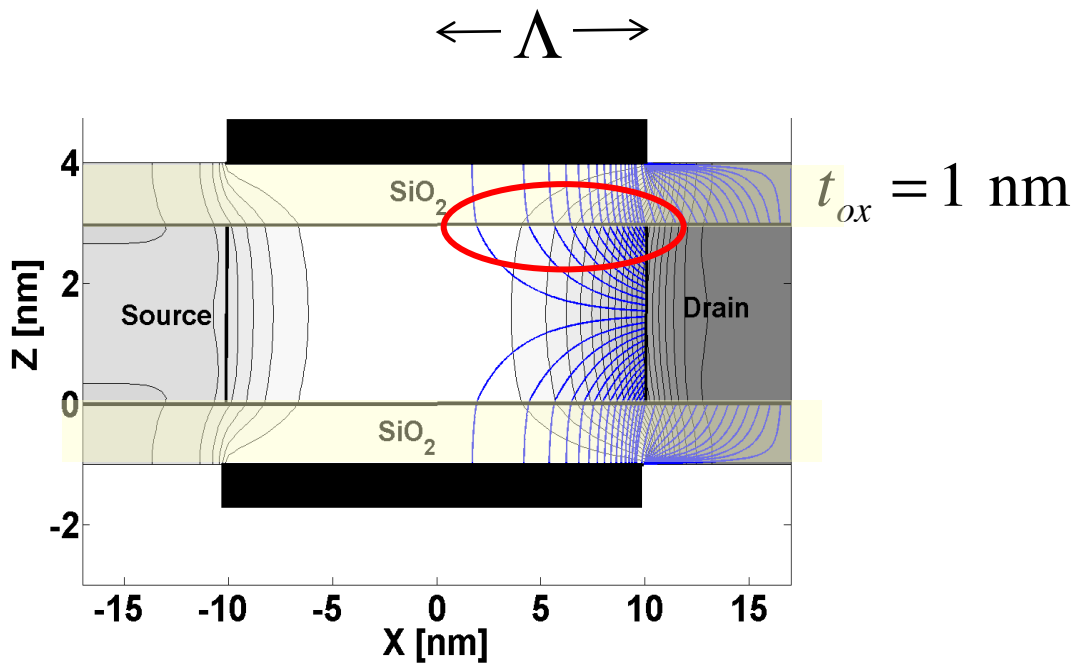
$I_D \sim Q_n(x=0) \times v_{inj}$



2D electrostatics

"Geometric screening"

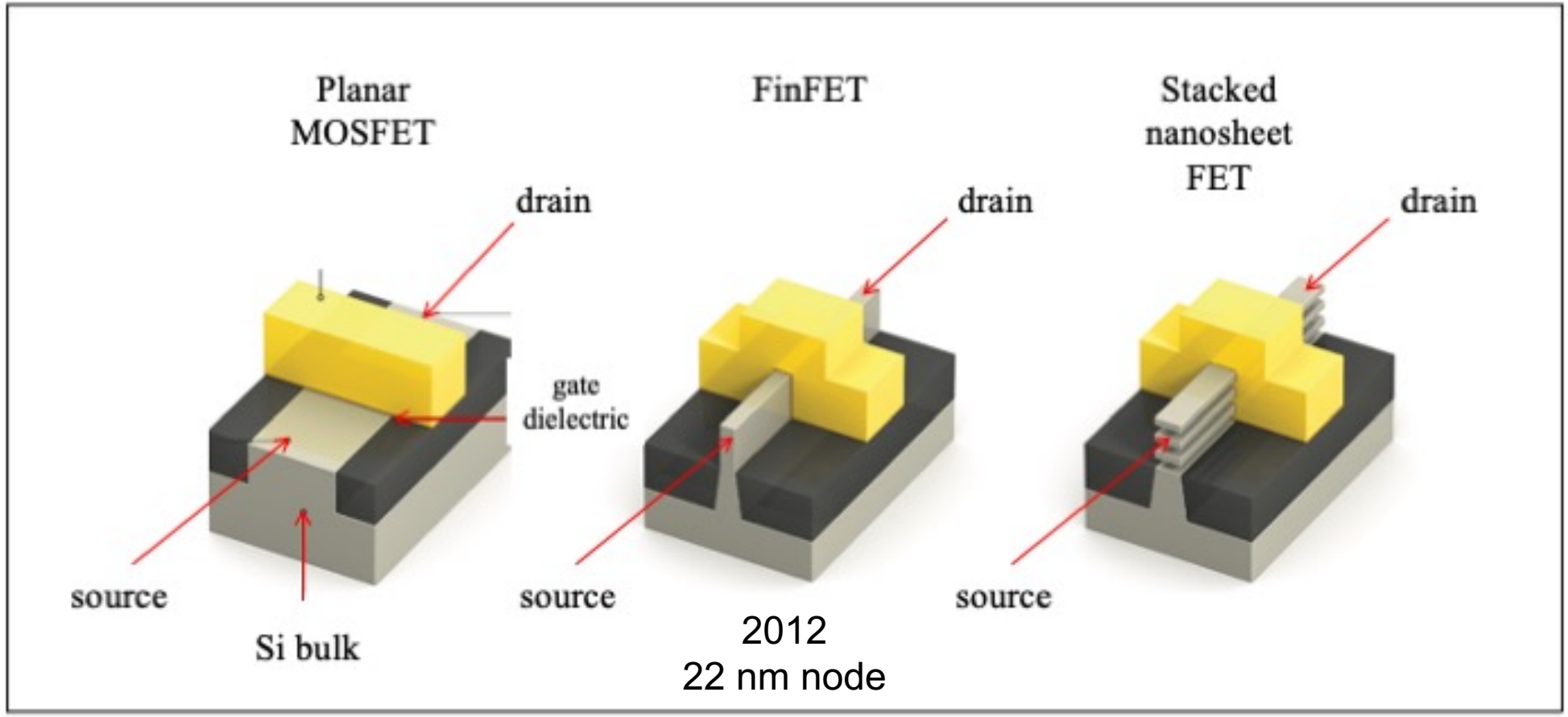




$t_{ox} = 1$ nm

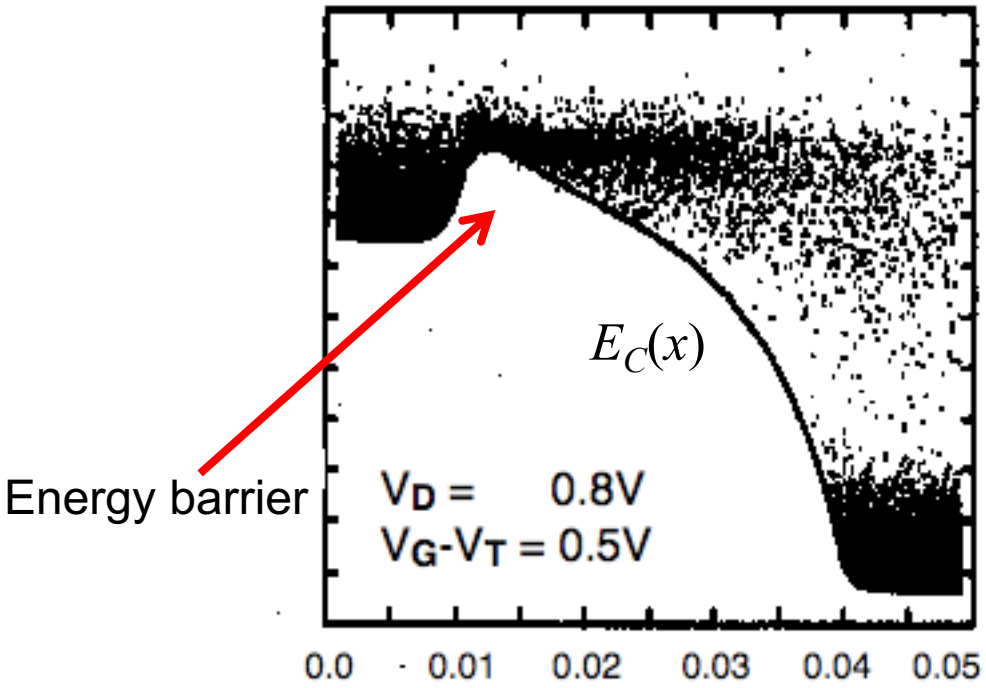
Off-state: $V_G = 0$ V, $V_D = 1$ V, $I_{off} = 0.1\mu A/\mu m$
(by H. Pal, Purdue, 2012)

Modern MOSFET structures



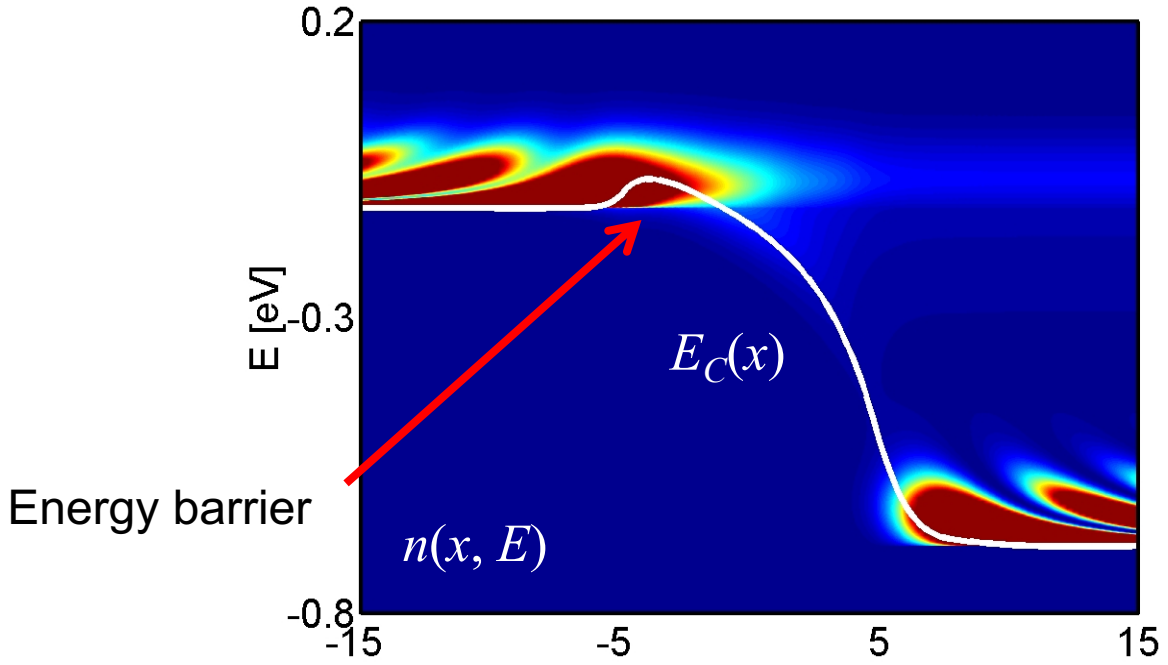
Electron transport in MOSFETs: Micro- to nano-scale

L = 30 nm



D. Frank, S. Laux, and M. Fischetti, Int. Electron Dev. Mtg., Dec., 1992.

L = 10 nm



Ren, Venugopal, Goasguen, Datta, and Lundstrom, TED, **50**, p. 2185, 2003.

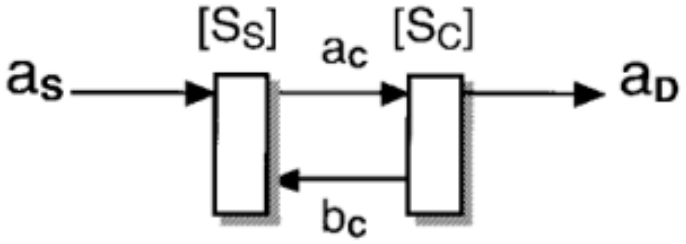
Essential physics of the Si MOSFET

IEEE ELECTRON DEVICE LETTERS, VOL. 18, NO. 7, JULY 1997

Elementary Scattering Theory of the Si MOSFET

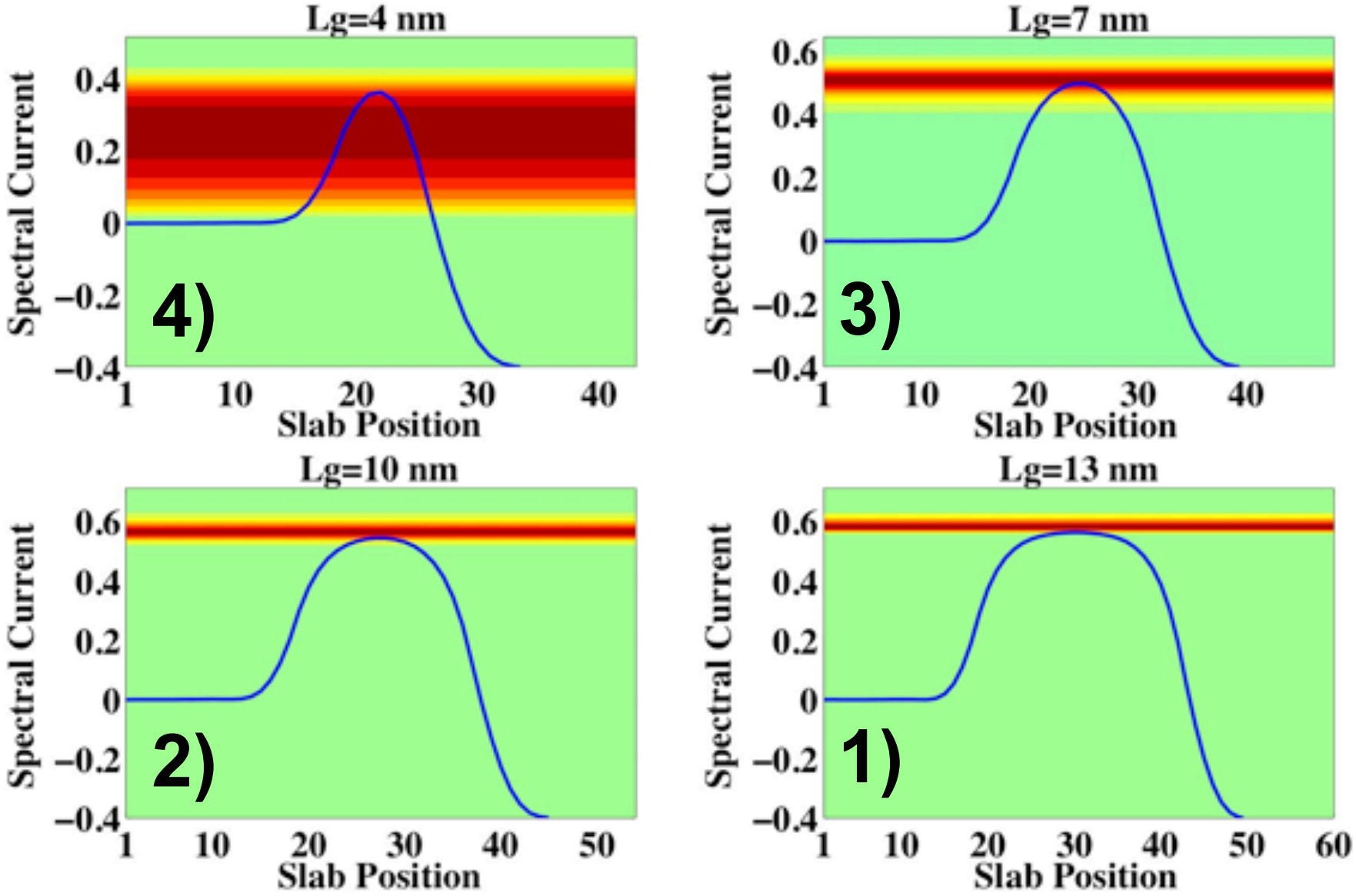
Mark Lundstrom

***Abstract*— A simple one-flux scattering theory of the silicon MOSFET is introduced. Current–voltage ($I-V$) characteristics are expressed in terms of scattering parameters rather than a mobility. For long-channel transistors, the results reduce to conventional drift-diffusion theory, but they also apply to devices in which the channel length is comparable to or even shorter than the mean-free-path. The results indicate that for very short channels the transconductance is limited by carrier injection from the source. The theory also indicates that evaluation of the drain current in short-channel MOSFET's is a near-equilibrium transport problem, even though the channel electric field is large in magnitude and varies rapidly in space.**



$$I_{Dsat} = C_{ox} W v_T \left(\frac{1-r_c}{1+r_c}\right)(V_{GS} - V_T)$$

The ultimate Si MOSFET



M. Luisier,
ETH Zurich /
Purdue

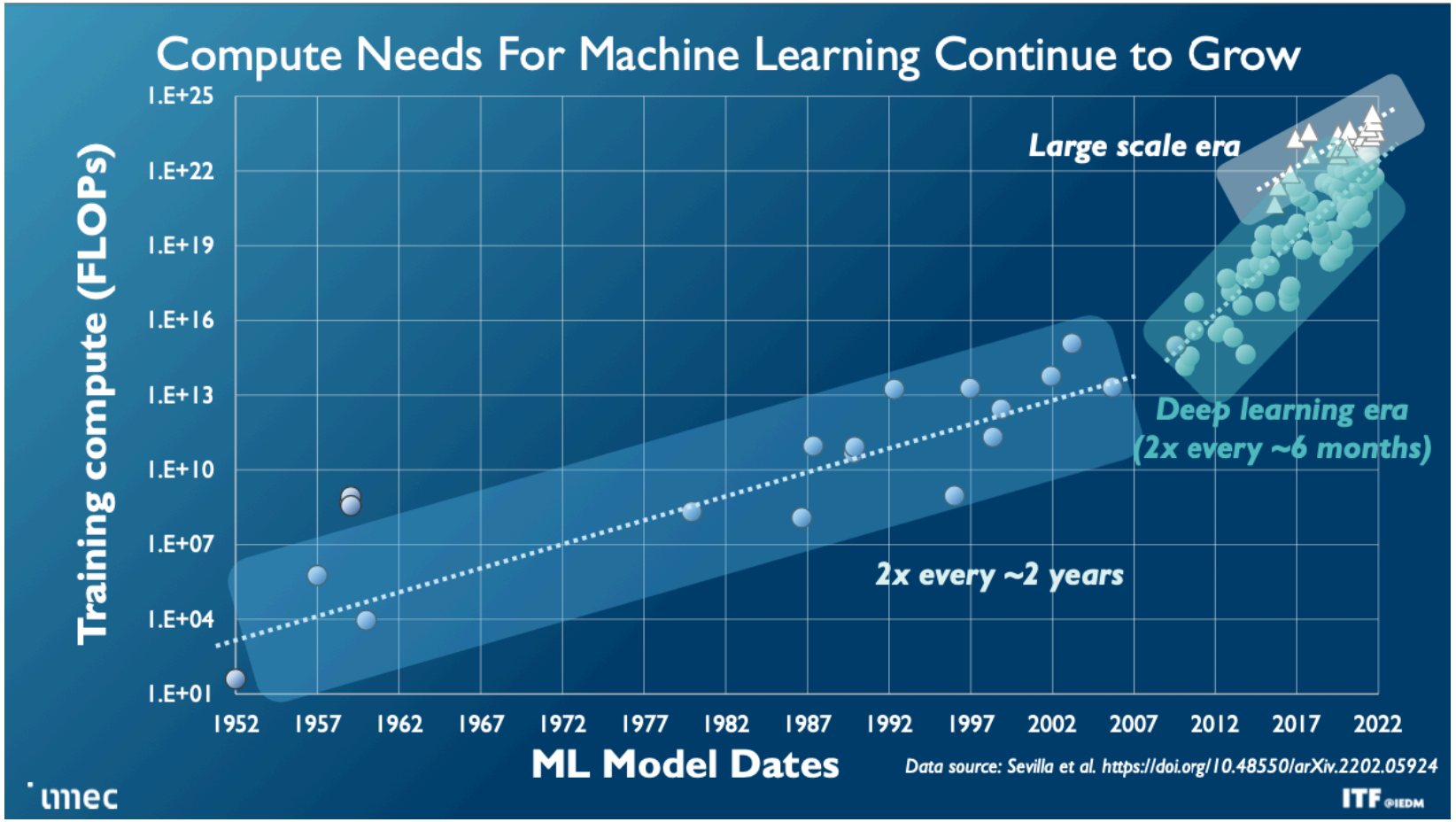
Part 2

Essential Physics of the Ultimate MOSFET

and

the next 20 years of semiconductor technology

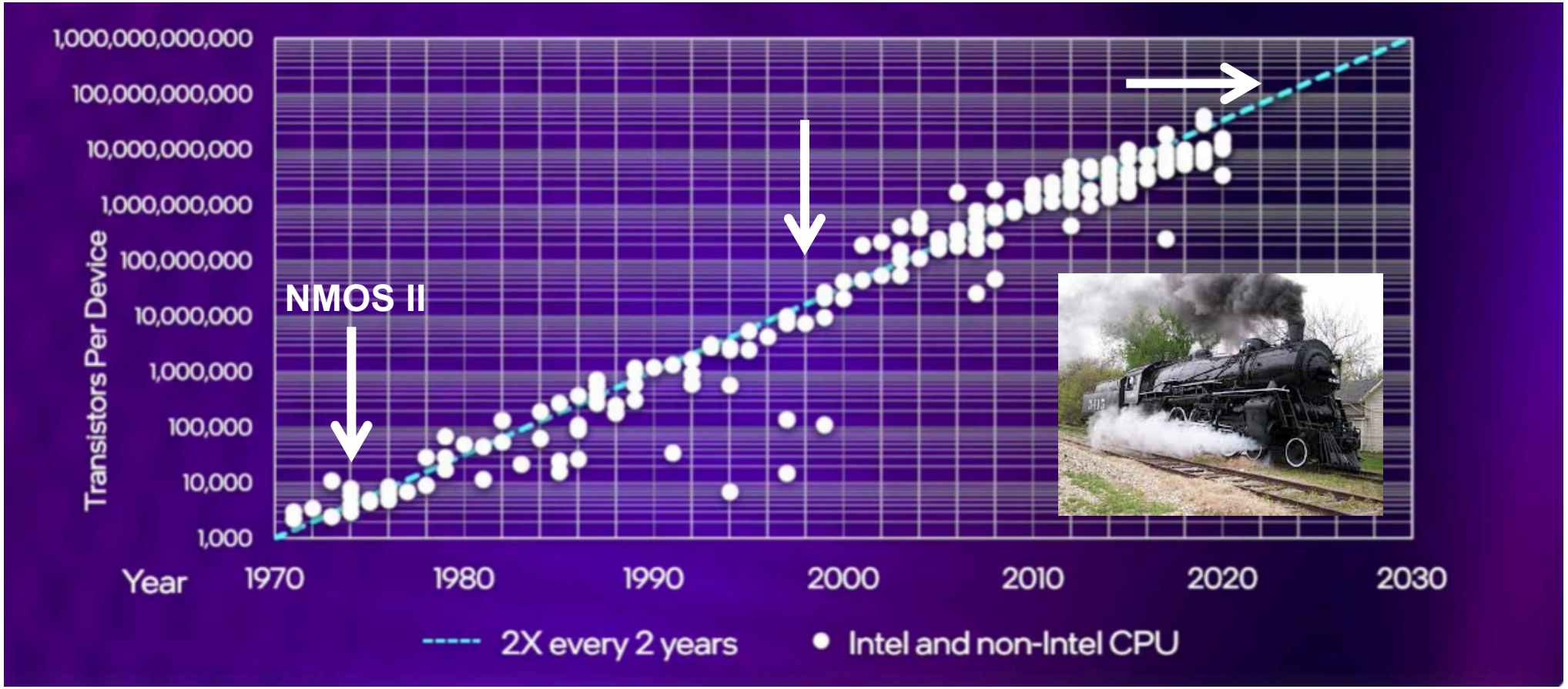
Exploding demand for computing and memory



The AI “gold rush”

Lundstrom: February 2024

Moore's Law



5000 nm

10,000 /chip

5 nm

>10,000,000,000 /chip

Summary

APPLIED PHYSICS

Moore's Law Forever?

Mark Lundstrom

When Gordon Moore predicted in 1965 that the number of transistors per integrated circuit chip would continue to double in each technology generation, there were just 30 transistors on a chip. Today, transistor counts—a measure of the capability of an electronic system—exceed a few hundred million for logic chips and even more for memory chips. How long can Moore's law continue?

Enhanced online at www.sciencemag.org/cgi/content/full/299/5604/210

The semiconductor industry follows Moore's law by shrinking transistor dimensions. But transistors cannot be scaled down infinitely. A few years ago, as critical dimensions approached 100 nm, a number of formidable challenges arose (*1*). It seemed that progress would slow, but during the past few years, device scaling has accelerated, as evidenced by several talks at the recent International Electron Devices Meeting (IEDM) (*2*).

The author is at the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA. E-mail: lundstro@purdue.edu

Today's electronic devices are based on the metal oxide semiconductor field-effect transistor (MOSFET), which consists of source and drain electrodes, through which current can flow, and a gate electrode, which controls the current through the other two (see the figure). MOSFETs operate on a simple principle: When the gate voltage is low, an energy barrier prevents electrons from flowing from source to drain, whereas a high gate voltage lowers the energy barrier, allowing current to flow (see the figure). The gate electrode is separated from the silicon channel by a thin insulating layer to prevent the flow of gate current.

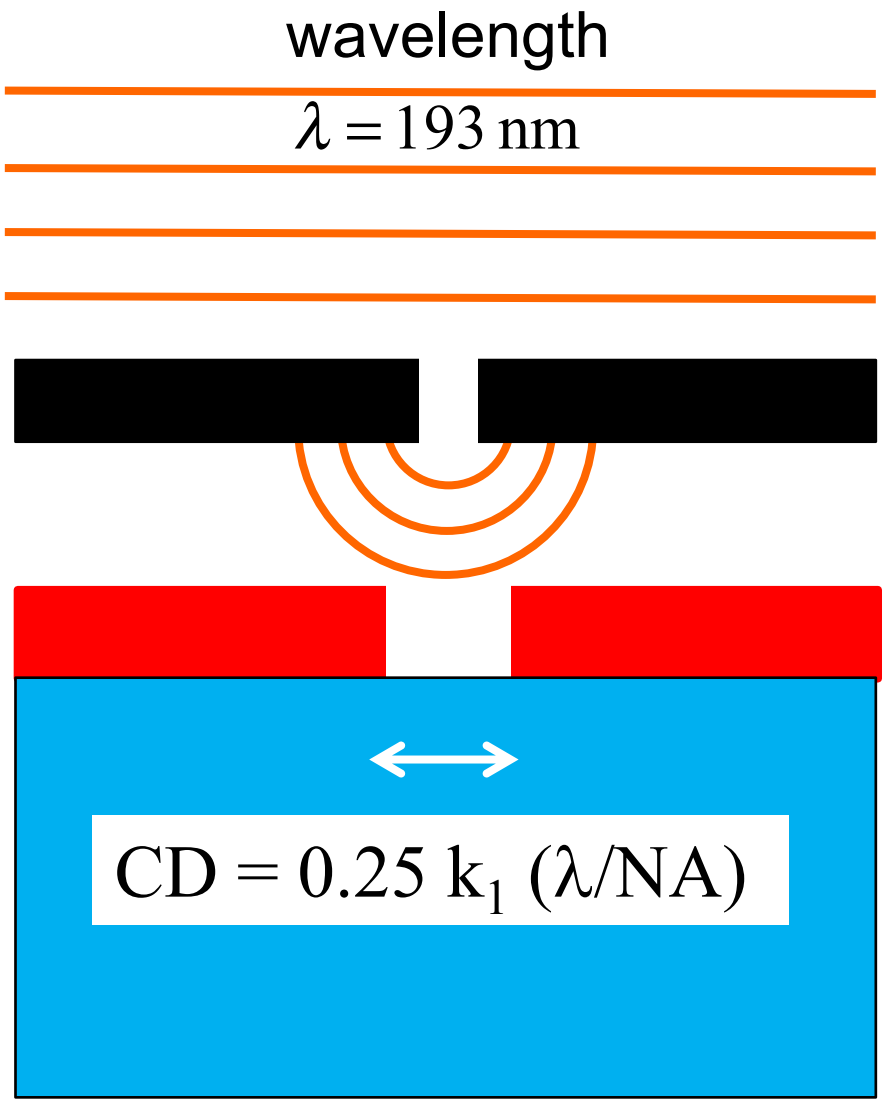
To comply with Moore's law, the transistor designer must shrink the distance between source and drain by a factor of $\sqrt{2}$ in each technology generation. This reduces the area by a factor of 2, thereby doubling the number of transistors per chip. Remarkable advances in subwavelength lithography allow current-generation technologies with gate lengths of 65 nm to be manufactured. Economic considerations have not yet slowed progress, and state-of-the-art technology still operates far below fundamental limits imposed by thermodynamics and quantum mechanics (*3*). The

Science, **299**, 210, 2003

"For the past 30 years, we have known what to do: **make transistors smaller.** Progress continues at a breathtaking pace, but transistor scaling is approaching its limit. When that limit is reached, **things must change,** but that does not mean that Moore's law has to end."

Making patterns smaller

90 nm
60 nm
45 nm
32 nm
22 nm
14 nm
10 nm
7 nm
5 nm
3 nm
2 nm



λ = 13.5 nm

ASML TWINSCAN NXT:2050i

Apple M2 Max

January 2023

67 billion transistors

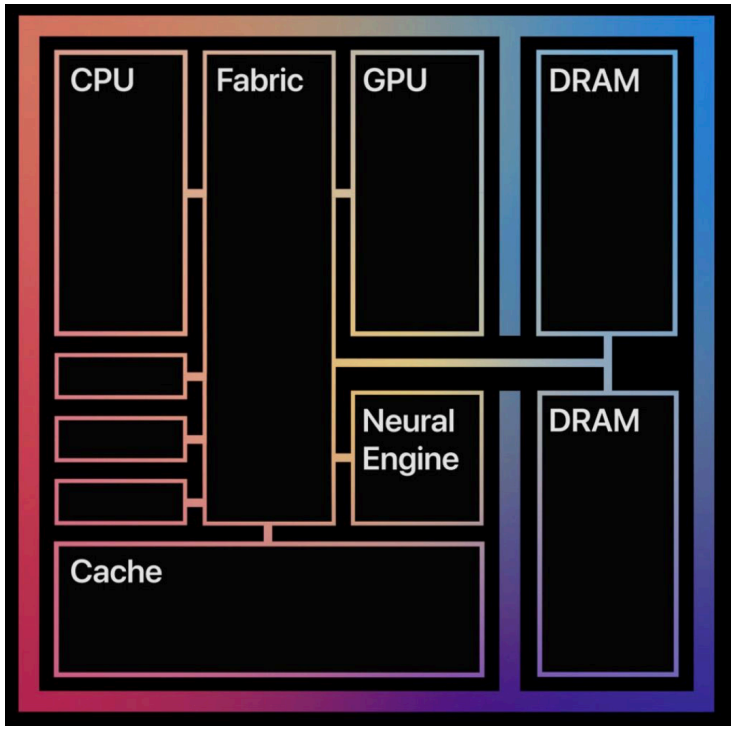
12 core CPU

38 core GPU

16 core neural engine

3.5 GHz clock

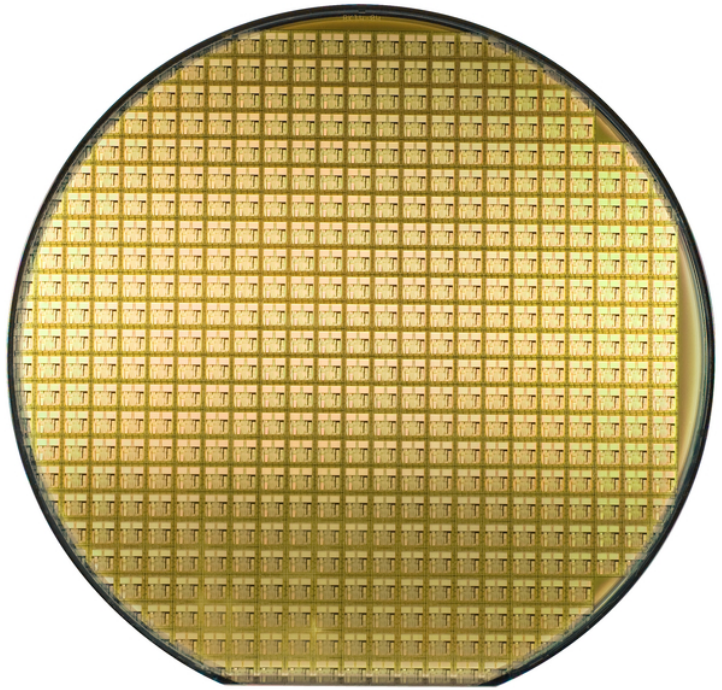
TSMC 5 nm process



System on a Chip (SoC)

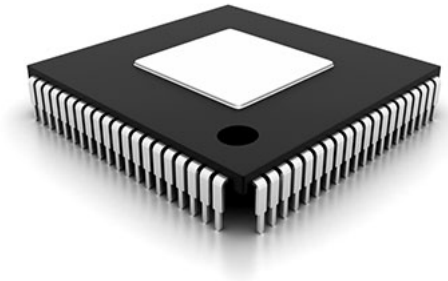
System on Chip → System in Package (SiP)

300 mm (12”) wafer



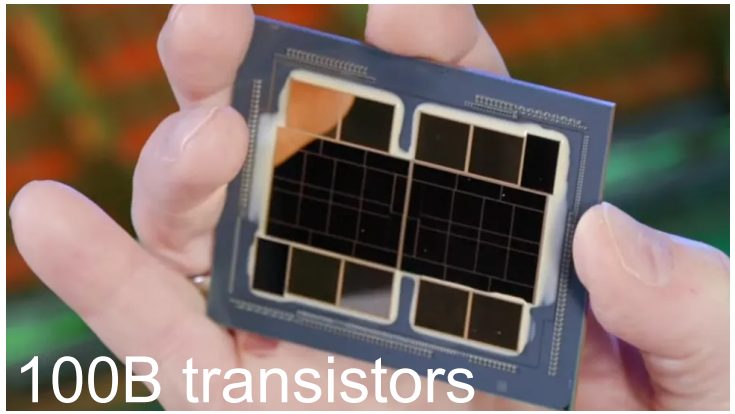
chip

package



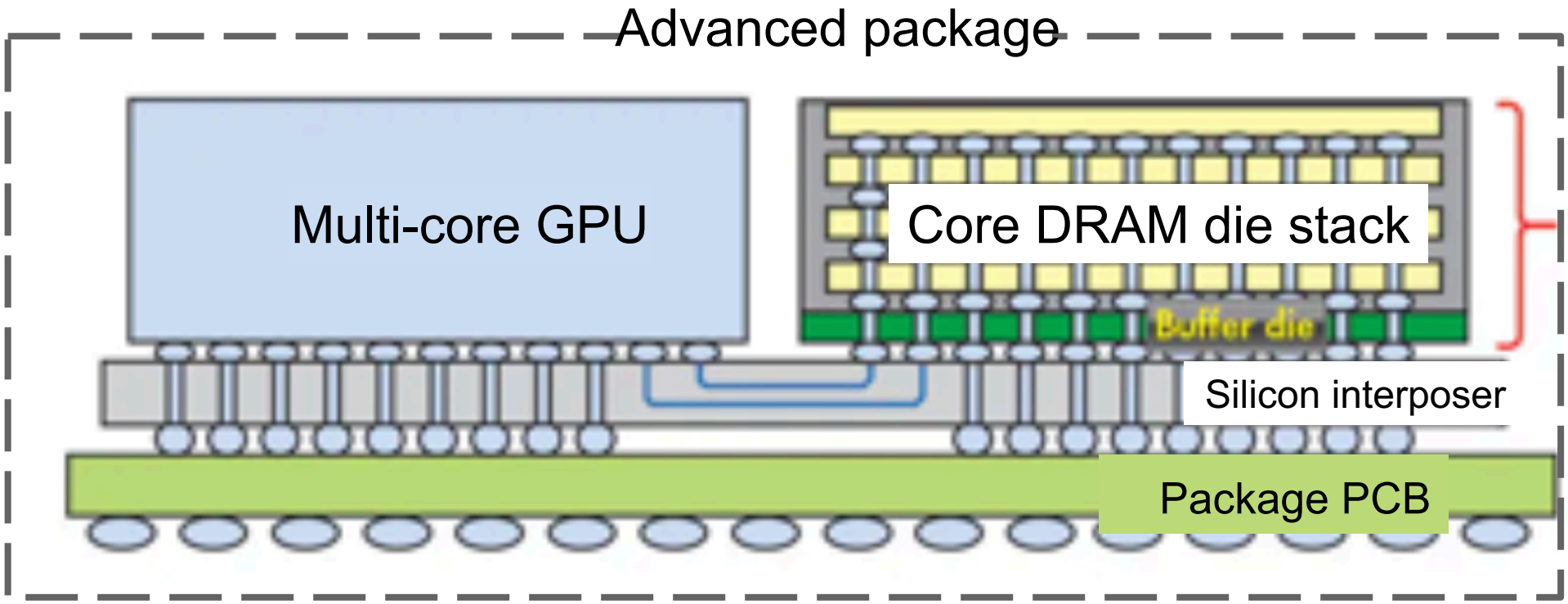
-one chip
-low tech
-cost is everything

advanced package



-complete electronic System in Package (SiP)

GPUs and HBM

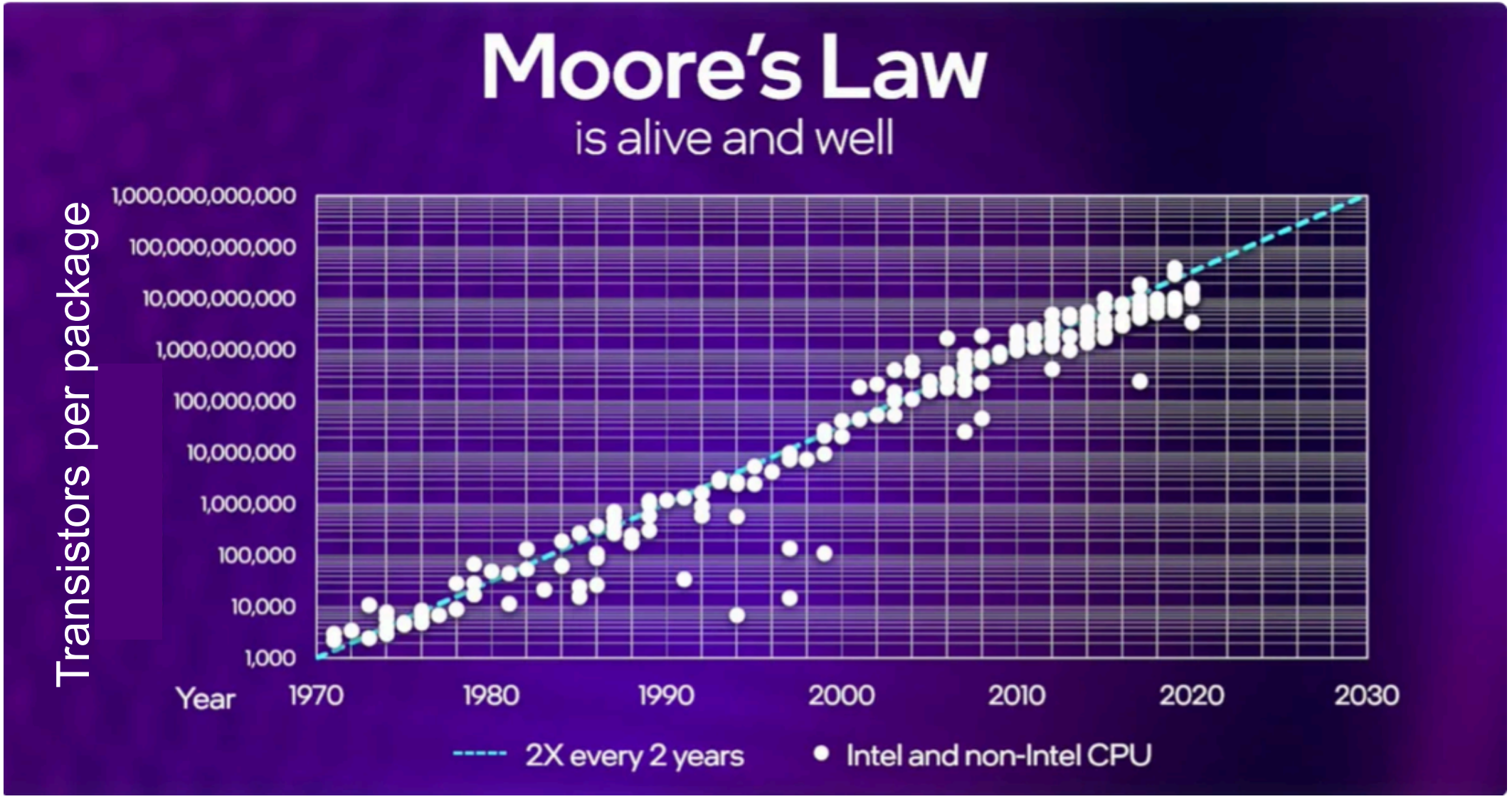


HBM
Stacked DRAM dies
Connected by Through Silicon Vias (TSVs)

16 layers in HBM3

High bandwidth memory closely connected through a Si interposer with a fast, multi-core GPU, can supply the CPUs with the data needed for LLMs.

Moore's Law Forever?



The next 20 years

APPLIED PHYSICS

Moore's Law Forever?

Mark Lundstrom

When Gordon Moore predicted in 1965 that the number of transistors per integrated circuit chip would continue to double in each technology generation, there were just 30 transistors on a chip. Today, transistor counts—a measure of the capability of an electronic system—exceed a few hundred million for logic chips and even more for memory chips. How long can Moore's law continue?

Enhanced online at www.sciencemag.org/cgi/content/full/299/5604/210

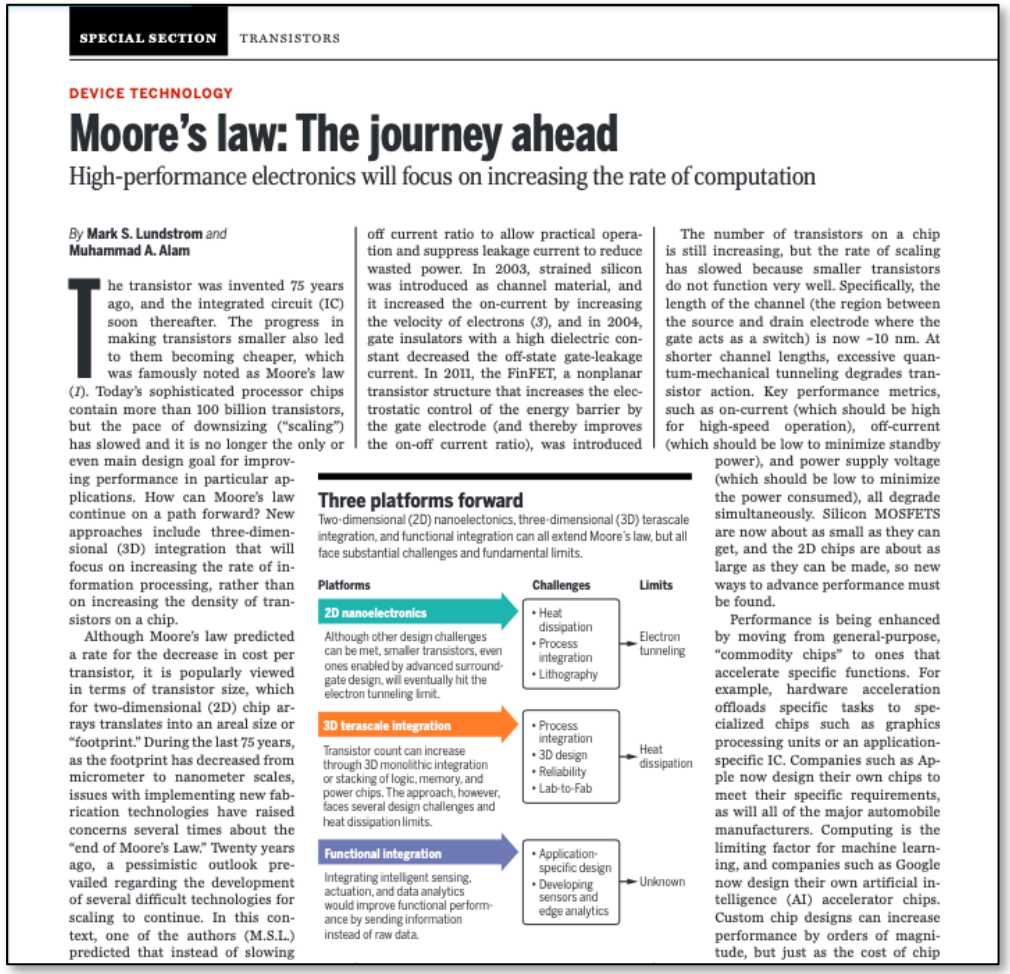
The semiconductor industry follows Moore's law by shrinking transistor dimensions. But transistors cannot be scaled down infinitely. A few years ago, as critical dimensions approached 100 nm, a number of formidable challenges arose (*1*). It seemed that progress would slow, but during the past few years, device scaling has accelerated, as evidenced by several talks at the recent International Electron Devices Meeting (IEDM) (*2*).

The author is at the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA. E-mail: lundstro@purdue.edu

Today's electronic devices are based on the metal oxide semiconductor field-effect transistor (MOSFET), which consists of source and drain electrodes, through which current can flow, and a gate electrode, which controls the current through the other two (see the figure). MOSFETs operate on a simple principle: When the gate voltage is low, an energy barrier prevents electrons from flowing from source to drain, whereas a high gate voltage lowers the energy barrier, allowing current to flow (see the figure). The gate electrode is separated from the silicon channel by a thin insulating layer to prevent the flow of gate current.

To comply with Moore's law, the transistor designer must shrink the distance between source and drain by a factor of $\sqrt{2}$ in each technology generation. This reduces the area by a factor of 2, thereby doubling the number of transistors per chip. Remarkable advances in subwavelength lithography allow current-generation technologies with gate lengths of 65 nm to be manufactured. Economic considerations have not yet slowed progress, and state-of-the-art technology still operates far below fundamental limits imposed by thermodynamics and quantum mechanics (*3*). The

Science, **299**, 210, 2003



SPECIAL SECTION TRANSISTORS

DEVICE TECHNOLOGY

Moore's law: The journey ahead

High-performance electronics will focus on increasing the rate of computation

By **Mark S. Lundstrom** and **Muhammad A. Alam**

The transistor was invented 75 years ago, and the integrated circuit (IC) soon thereafter. The progress in making transistors smaller also led to them becoming cheaper, which was famously noted as Moore's law (*1*). Today's sophisticated processor chips contain more than 100 billion transistors, but the pace of downsizing ("scaling") has slowed and it is no longer the only or even main design goal for improving performance in particular applications. How can Moore's law continue on a path forward? New approaches include three-dimensional (3D) integration that will focus on increasing the rate of information processing, rather than on increasing the density of transistors on a chip.

Although Moore's law predicted a rate for the decrease in cost per transistor, it is popularly viewed in terms of transistor size, which for two-dimensional (2D) chip arrays translates into an areal size or "footprint." During the last 75 years, as the footprint has decreased from micrometer to nanometer scales, issues with implementing new fabrication technologies have raised concerns several times about the "end of Moore's Law." Twenty years ago, a pessimistic outlook prevailed regarding the development of several difficult technologies for scaling to continue. In this context, one of the authors (M.S.L.) predicted that instead of slowing off current ratio to allow practical operation and suppress leakage current to reduce wasted power. In 2003, strained silicon was introduced as channel material, and it increased the on-current by increasing the velocity of electrons (*3*), and in 2004, gate insulators with a high dielectric constant decreased the off-state gate-leakage current. In 2011, the FinFET, a nonplanar transistor structure that increases the electrostatic control of the energy barrier by the gate electrode (and thereby improves the on-off current ratio), was introduced

The number of transistors on a chip is still increasing, but the rate of scaling has slowed because smaller transistors do not function very well. Specifically, the length of the channel (the region between the source and drain electrode where the gate acts as a switch) is now ~10 nm. At shorter channel lengths, excessive quantum-mechanical tunneling degrades transistor action. Key performance metrics, such as on-current (which should be high for high-speed operation), off-current (which should be low to minimize standby power), and power supply voltage (which should be low to minimize the power consumed), all degrade simultaneously. Silicon MOSFETS are now about as small as they can get, and the 2D chips are about as large as they can be made, so new ways to advance performance must be found.

Performance is being enhanced by moving from general-purpose, "commodity chips" to ones that accelerate specific functions. For example, hardware acceleration offloads specific tasks to specialized chips such as graphics processing units or an application-specific IC. Companies such as Apple now design their own chips to meet their specific requirements, as will all of the major automobile manufacturers. Computing is the limiting factor for machine learning, and companies such as Google now design their own artificial intelligence (AI) accelerator chips. Custom chip designs can increase performance by orders of magnitude, but just as the cost of chip

Three platforms forward

Two-dimensional (2D) nanoelectonics, three-dimensional (3D) terascale integration, and functional integration can all extend Moore's law, but all face substantial challenges and fundamental limits.

Platforms	Challenges	Limits
2D nanoelectronics Although other design challenges can be met, smaller transistors, even ones enabled by advanced surround-gate design, will eventually hit the electron tunneling limit.	• Heat dissipation • Process integration • Lithography	Electron tunneling
3D terascale integration Transistor count can increase through 3D monolithic integration or stacking of logic, memory, and power chips. The approach, however, faces several design challenges and heat dissipation limits.	• Process integration • 3D design • Reliability • Lab-to-Fab	Heat dissipation
Functional integration Integrating intelligent sensing, actuation, and data analytics would improve functional performance by sending information instead of raw data.	• Application-specific design • Developing sensors and edge analytics	Unknown

Science, **378**, 722, 2022

CHIPS for America

$39 billion
for incentives

Two component programs to:

1. Attract large-scale investments in advanced technologies such as leading-edge logic and memory, and advanced packaging
2. Incentivize expansion of manufacturing capacity for mature and other types of semiconductors

$11 billion
for R&D

Four integrated programs to:

1. Conduct research and prototyping of advanced semiconductor technology
2. Strengthen semiconductor advanced packaging, assembly, and test
3. Enable advances in measurement science, standards, material characterization, instrumentation, testing, and manufacturing

$2B NSTC and $3B NAPMP

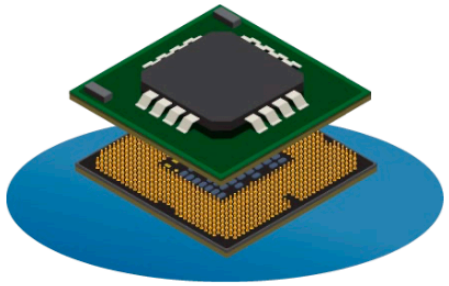
Plus CHIPS initiatives from other agencies, including DOD, State, NSF, and Treasury

DoD ME Commons
DARPA NGMM

CHIPS R&D Programs

Metrology

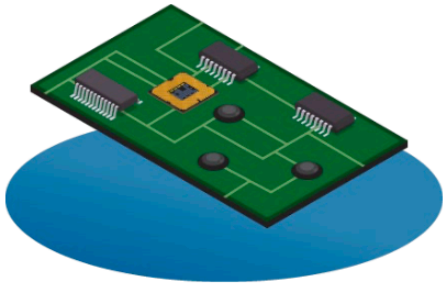
NIST



$2B

National Semiconductor Technology Center

300 mm Baseline CMOS R&D
Tech Centers on 10 topics



$3B

National Advanced Packaging Manufacturing Program

Adv. Package Pilot Facility
Tech Centers on 6 topics

Manufacturing USA

$200M

Manufacturing USA Institute

Digital Twins

Workforce Initiatives

Semiconductors@Purdue

1) NAPMP Technical Center

2) NSTC Technical Center

imec
@Purdue

ASIP
ATALLA INSTITUTE FOR
ADVANCED SYSTEM
INTEGRATION AND PACKAGING

Birck Nanotechnology Center
(Birck 2.0)

CSME
CENTER FOR SECURE
MICROELECTRONICS ECOSYSTEM

Purdue@Crane

Institute of Hard AI

(DARPA NGMM)

3) NSTC WFD Center

4) MUSA Institute Digital Twins

Summary

- Semiconductor technology is foundational and will continue to be.
- Simple ideas are powerful (e.g. zeros and ones and MOSFETs).
- Sustained incremental progress can be transformative.
- Universities can play an important role.