The physical basis of digital computing

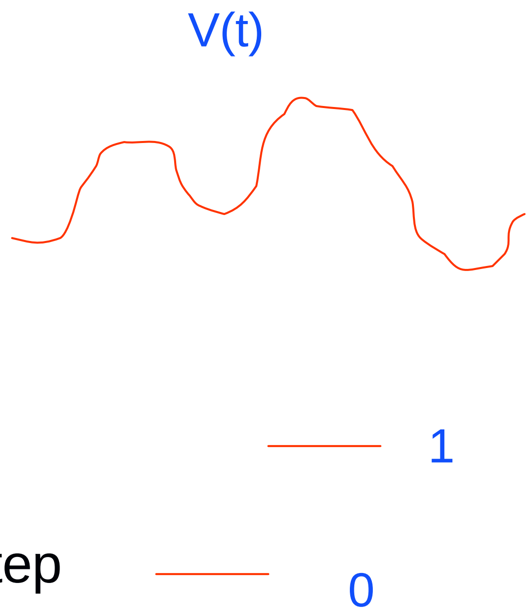
Henk van Houten

Philips Research Laboratories
5656 AA Eindhoven, The Netherlands
e-mail: henk.van.houten@philips.com

Contents

- Digital switching and the thermal limit
  - restoring logic
  - the MOSFET and Moore's law
  - dissipation and the discarding of information
  - adiabatic or reversible computing
- Digital switching and the quantum limit
  - Quantum ballistic transport
  - Single electron tunneling
  - The ultimate MOSFET
  - Quantum limit of switching devices
  - Applications of quantum devices?
- A hierarchy of limits (Meindl's classification)
- Lithography and the end of Moore's law
- Conclusions and References

Digital signal representation



- An analogue signal can represent many bits
  - but signal distortion is inevitable in complex systems
- A digital signal represents only a single bit (0,1)
  - because no system is perfect (noise, distortion)
  - simple standardization of signal levels
  - logic level restoration possible at each computational step

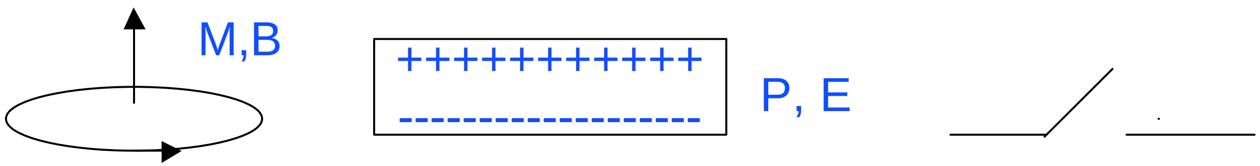
Indefinite extension possible
without error propagation

Digital computing has a physical basis

- computing is a physical process
  - logic devices have a finite physical extent
  - require a minimum time to perform their function
  - dissipate a switching energy

thermodynamics
electromagnetism
quantum mechanics
& information theory

- The physical basis sets the scale for size, speed, and power requirements of a computing system
- Many physical implementations are possible



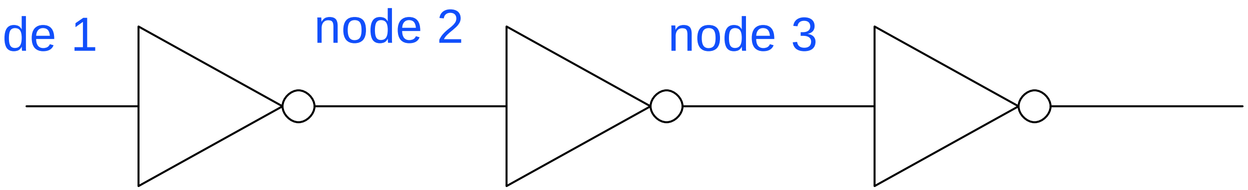
Spin, magnetization, field direction, polarization, mechanical switch
MOSFET channel conductance,...

Restoring logic devices must have gain

A digital signal is stored as a *signal energy*

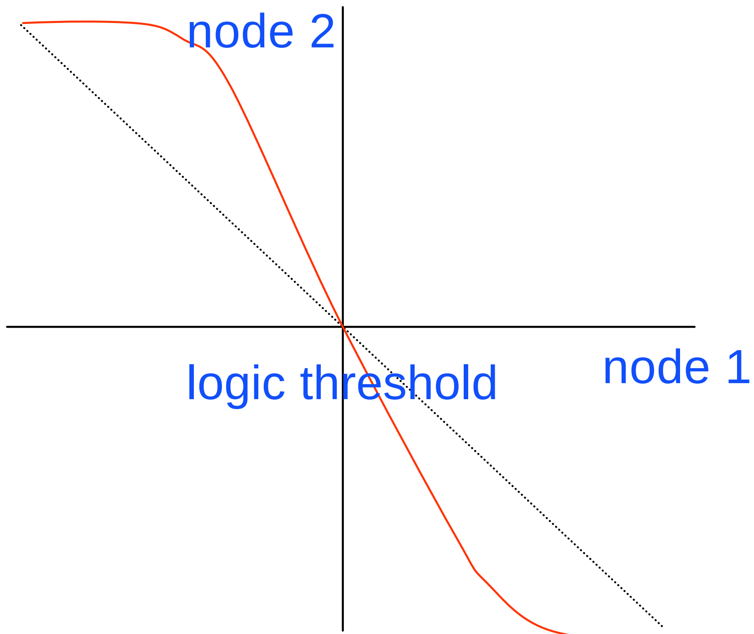
A logic circuit must be able to drive similar circuits

Inverter chain



Logic transfer characteristic
in valid range for “0” or “1”
the slope must be *less* than 1

so: around logic threshold the
slope must be *larger* than 1

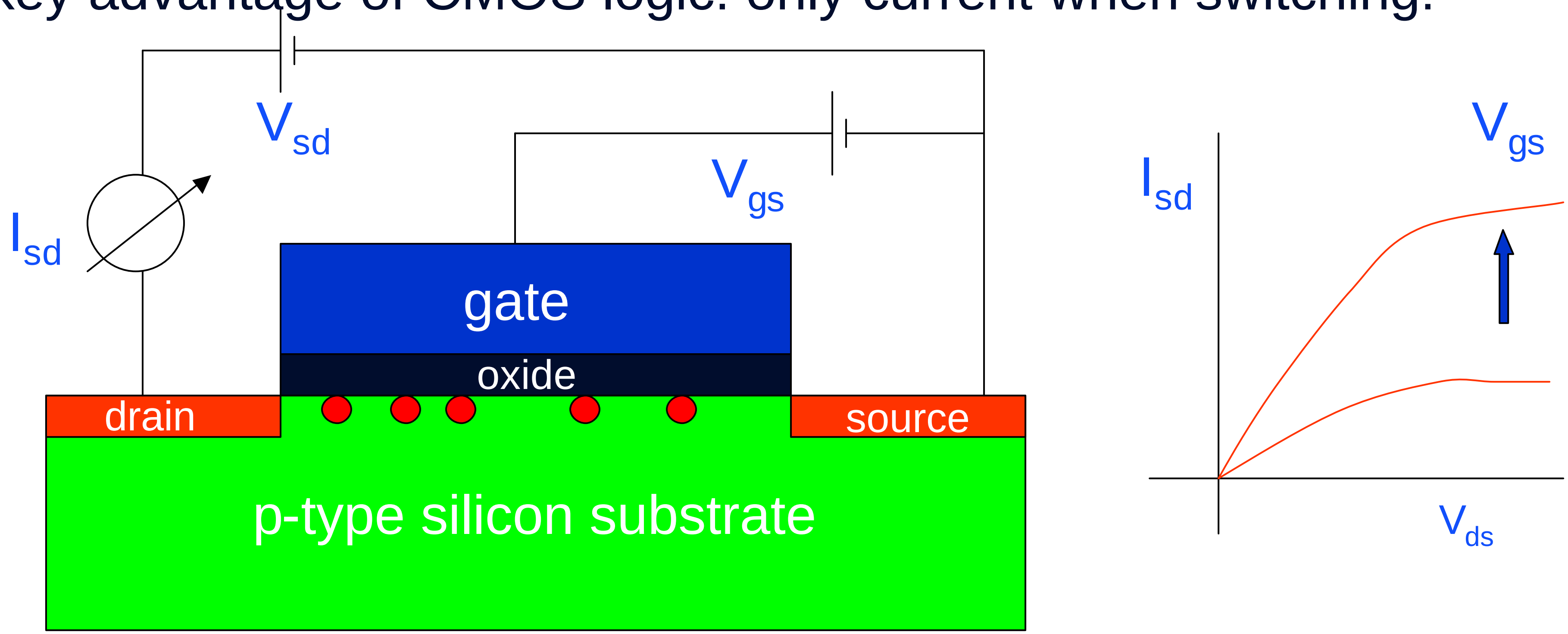


Circuit must have energy gain
Switch in finite time: power gain

Computer must have a power supply
separate from the signal path

The MOSFET

- A MOSFET is basically a switchable resistor with gain
- the charge in the channel is determined by the gate voltage
- key advantage of CMOS logic: only current when switching!



Induced charge density in the n-channel $en_{induced} = C(V_{gs}-V_t)$

Switching time

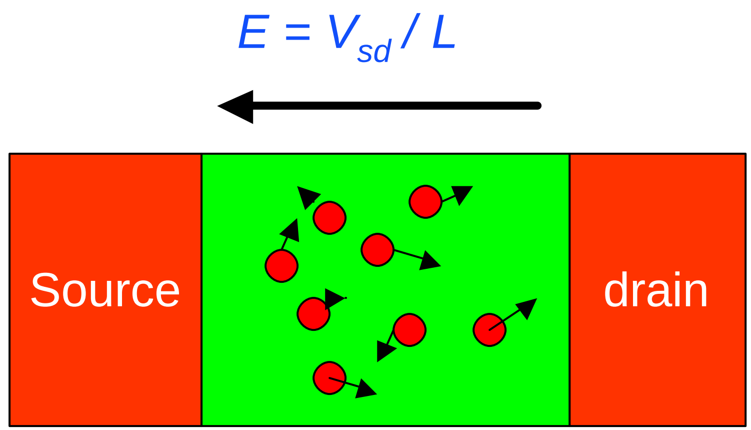
Current through a MOSFET

(small source-drain voltage)

$$I = e\, n_{induced}\, v_{drift}$$

Transit time

$$\tau = L/v_{drift} = L^2/\mu V_{ds}$$



Charge density $e n_{induced} = C(V_{gs}-V_t)$

mobility μ

drift velocity $v_{drift} = \mu E$

Saturation occurs when $V_{ds} \geq V_g - V_t$

Beyond this point, the transit time no longer decreases.

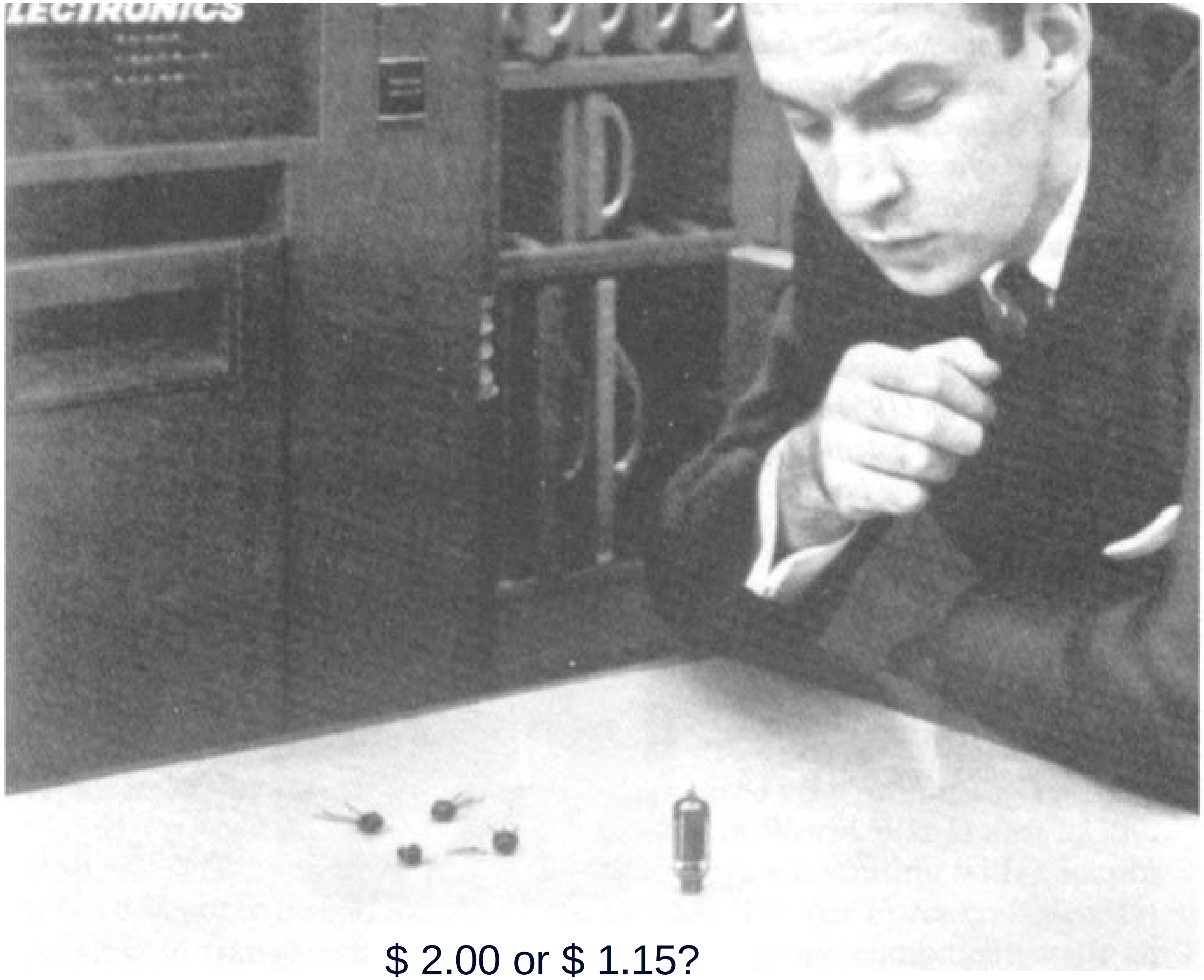
Switching time is essentially the charging time of the next gate

$$R\,C \approx \tau$$

The transit time is the basic measure of switching delay

A worried manager: miniaturization

Advertisement of General Electric in Scientific American 1961



$ 2.00 or $ 1.15?

Moore's law: the number of transistors on a chip doubles every 12 (18) months

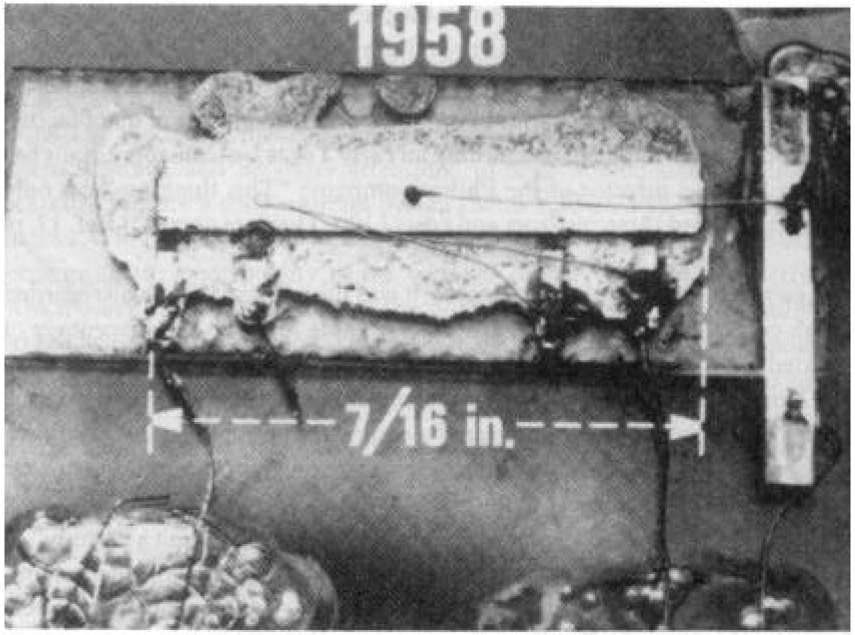
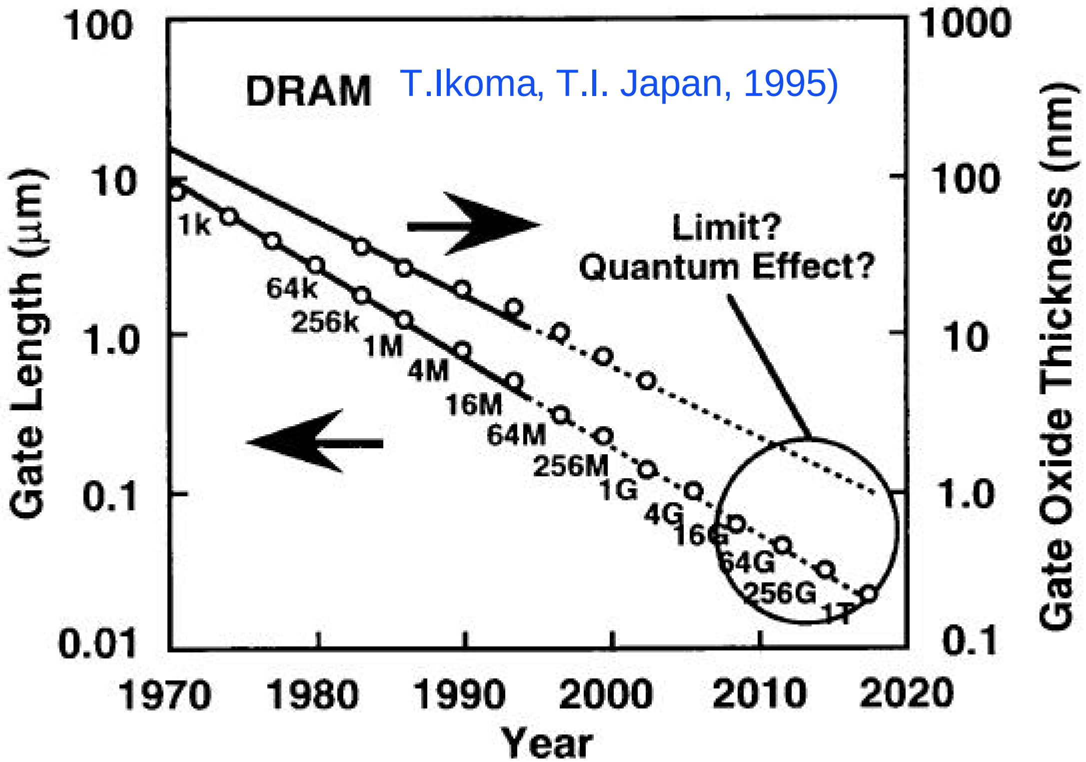


FIG. 64. The first integrated circuit, built by Texas Instruments on the basis of Kilby's sketch. ($\frac{7}{16}$ in. equals about 11 mm).



in10 years $\alpha \approx 5$ (Moore's law)

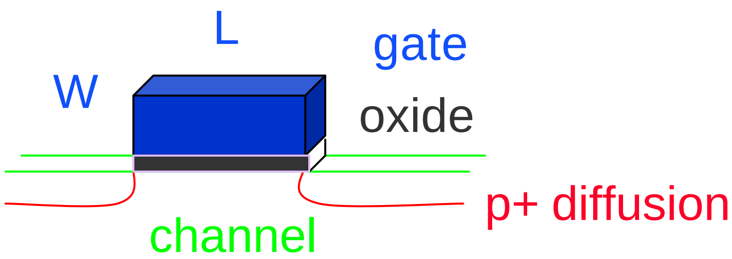
$$L \Rightarrow L/\alpha$$

$$d_{oxide} \Rightarrow d_{oxide} / \alpha$$

Henk van Houten, LATSIS symposium, June 2000

Scaling of a MOSFET

Scale all geometrical dimensions down by α
and keep the electric fields constant



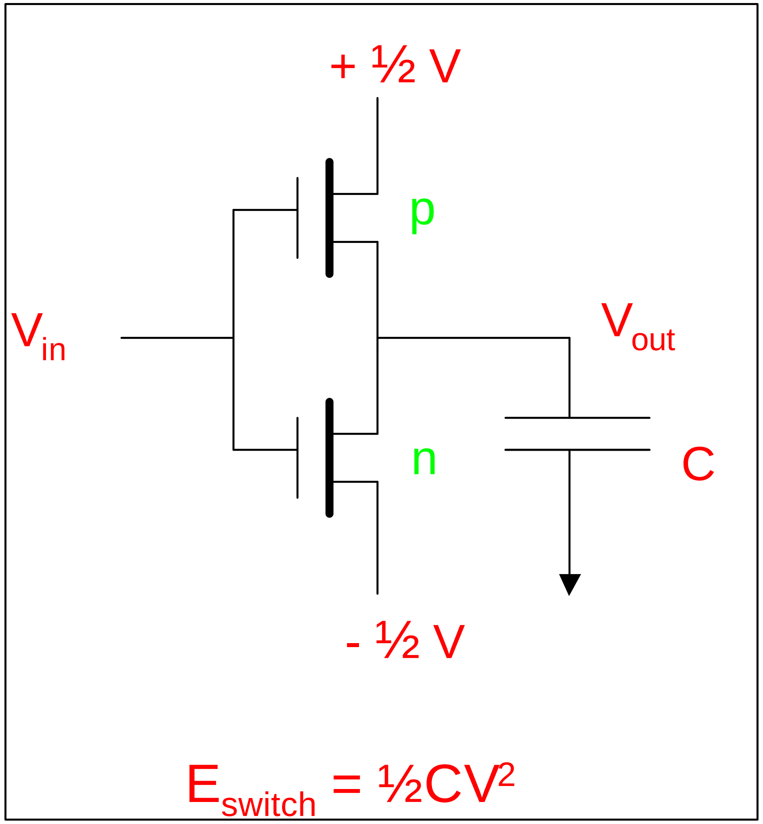
L	$\Rightarrow$	L/α
W	$\Rightarrow$	W/α
d_{oxide}	$\Rightarrow$	d_{oxide} / α
V	$\Rightarrow$	V/ α

gate capacitance	$C = \varepsilon W L / d_{oxide}$	$\propto 1/\alpha$
switching energy	$E = ½ C V^2$	$\propto 1/\alpha^3$
switching time	$\tau \propto L^2/V$	$\propto 1/\alpha$
switching power	$P = E/\tau$	$\propto 1/\alpha^2$

Power per unit area remains constant

Henk van Houten, LATSIS symposium, June 2000

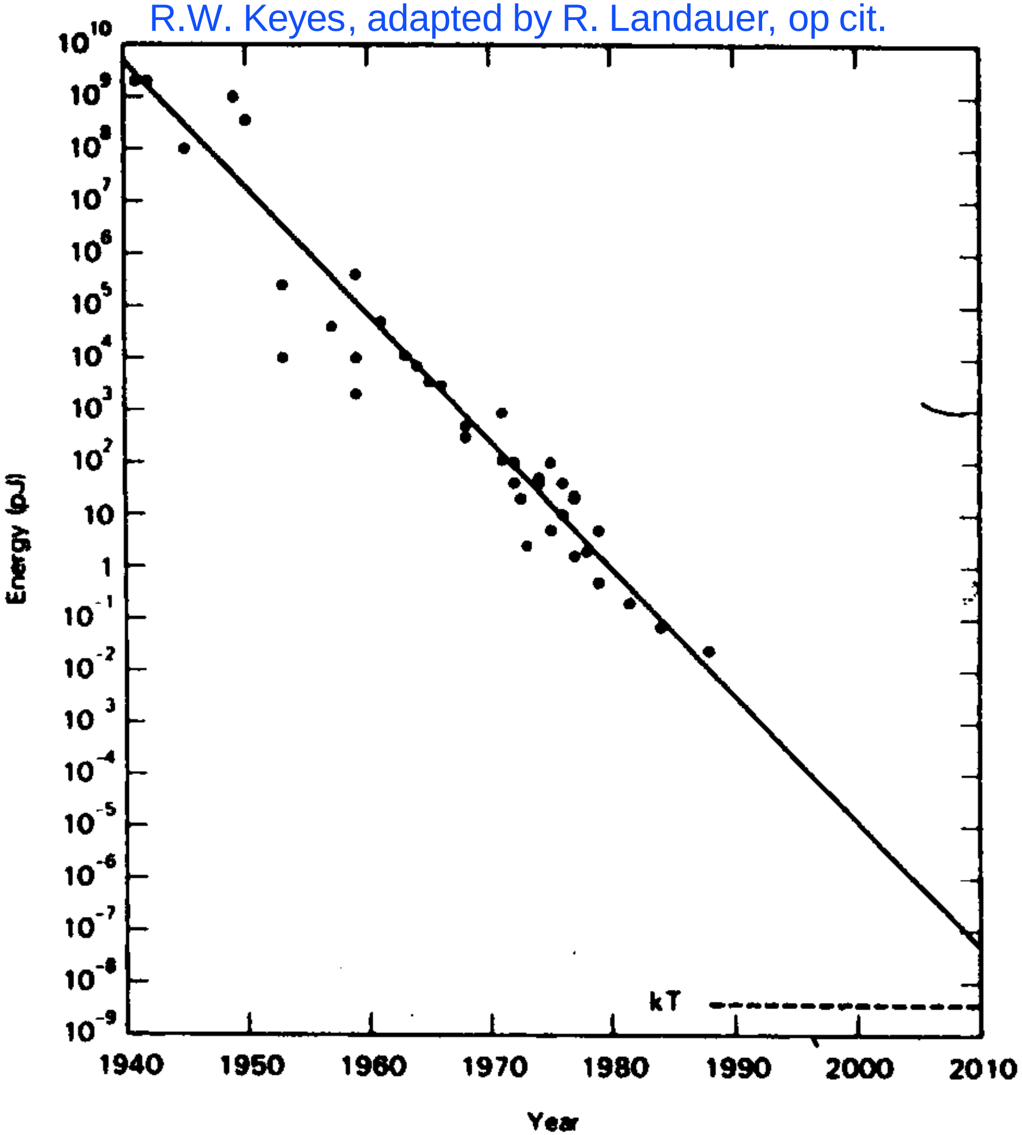
Scaling of switching energy



Scaling : $E = ½ C V^2 \propto 1/\alpha^3$

in10 years $\alpha \approx 5$ (Moore's law)

R.W. Keyes, adapted by R. Landauer, op cit.



Henk van Houten, LATSIS symposium, June 2000

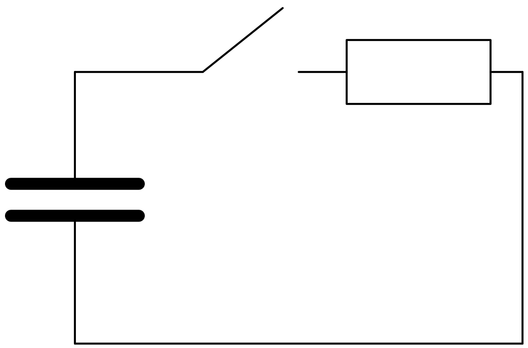
Quasi-adiabatic computing

Can the switching energy be reduced below $\frac{1}{2} C V^2$?

Yes, using “adiabatic” logic.

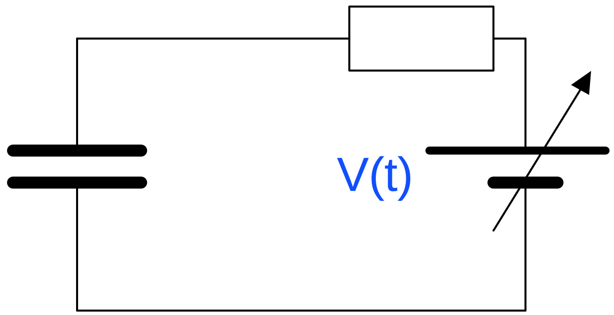
Illustration of quasi-adiabatic discharging

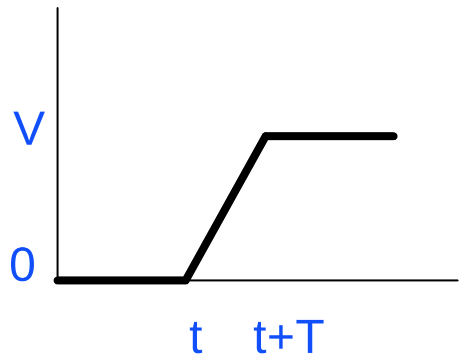
Discharging over a resistance



Energy dissipated is $\frac{1}{2} C V^2$

Discharging at constant current (ramped power supply)



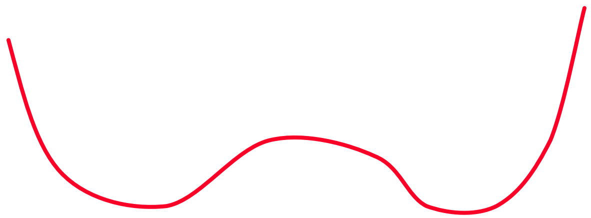


Energy dissipated is $\frac{1}{2} C V^2 (2RC/T)$

Power dissipation may be reduced at the cost of switching speed
The energy saved is stored in the power supply.
but: reduction of supply voltage has similar effect (and seems more practical)

How small can the dissipation be?

Why does the energy of a logic gate (and therefore of a digital computer) have to depend on its logical state?



Because otherwise there is no force to drive the transition, or to maintain the stable state!

Classical physics: thermal equilibrium noise 4kTRB induces transitions; switching energy should exceed kT

To drive a (number of) transitions in a deterministic and irreversible fashion, energy *must* be dissipated in the environment to which the system is coupled

Side step: Communication

- Shannon: a minimal energy is associated with the transmission of a bit over a perfect channel in the presence of (white Gaussian additive) noise

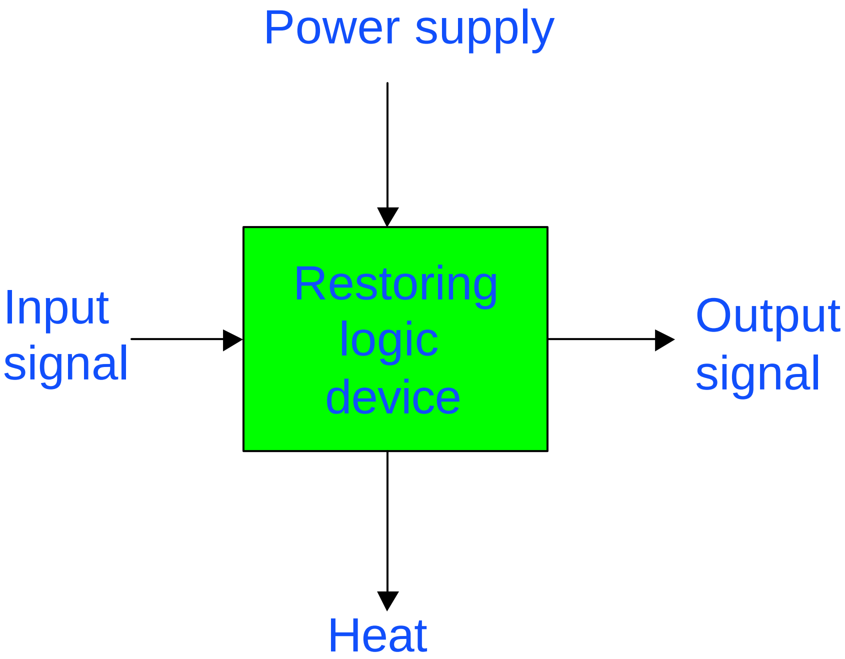
$$C = B \ln [(P+N)/N)]$$

C	channel capacity
P	received power
B	bandwidth
$N = kT\,B$	noise power

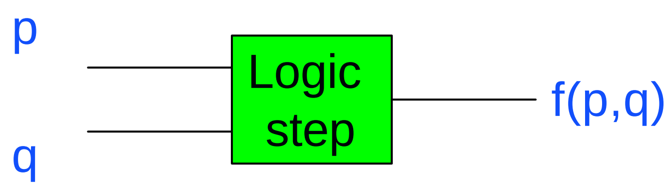
- P/C has a minimum value of kT ln 2

Landauer:
note that it is not clear that this energy has to be dissipated
note that this is an analysis of a specific case

The second law of thermodynamics and the arrow of time



Digital computing typically involves discarding of information (Landauer)



p, q can not be recovered uniquely from the non-linear function f(p,q)

In a statistical sense, an *irreversible* computational process can only be driven forward deterministically at the cost of a *minimal* energy dissipation of $k_B T \ln 2$ per erased bit (entropy drop of k_B ln2 per erased bit)

Reversible computers

- Information does not have to be discarded: use a series of 1: 1 mappings
- This is an extension of the “quasi-adiabatic” principle discussed before to the fully adiabatic case
- classical *reversible* circuit architectures have been proposed (Bennett, Fredkin, …)
- the price to pay: speed, HW complexity, the system has to be flawless, ...
- This has been the basis for the field of quantum computing, see next lecture by Gianni Blatter

Is a switching energy of kT enough?

probability of error: Boltzmann factor

$$P \approx \exp(-E_{switching}/kT)$$

$$E_{switching} > kT \ln(\text{ mean time between error}/\ \tau)$$

practical limit probably closer to 100 kT

(other error sources: cosmic radiation, synchronization error, etc)

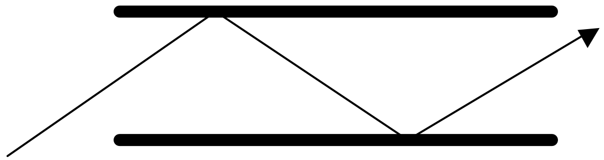
Contents

- Digital switching and the thermal limit
- **Digital switching and the quantum limit**
- A hierarchy of limits (Meindl's classification)

Quantum ballistic transport

Ballistic transport occurs on length scales short compared to the mean free path

Specular reflection off the boundaries of a conducting channel

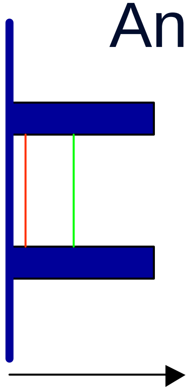


Infinite conductance?

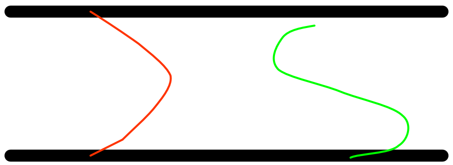
The De Broglie *wavelength* of a conduction electron

$\lambda = h / m v_F$ Typically 50 nm in a MOSFET or GaAs FET

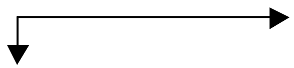
An electron waveguide has 1-dimensional subbands



energy



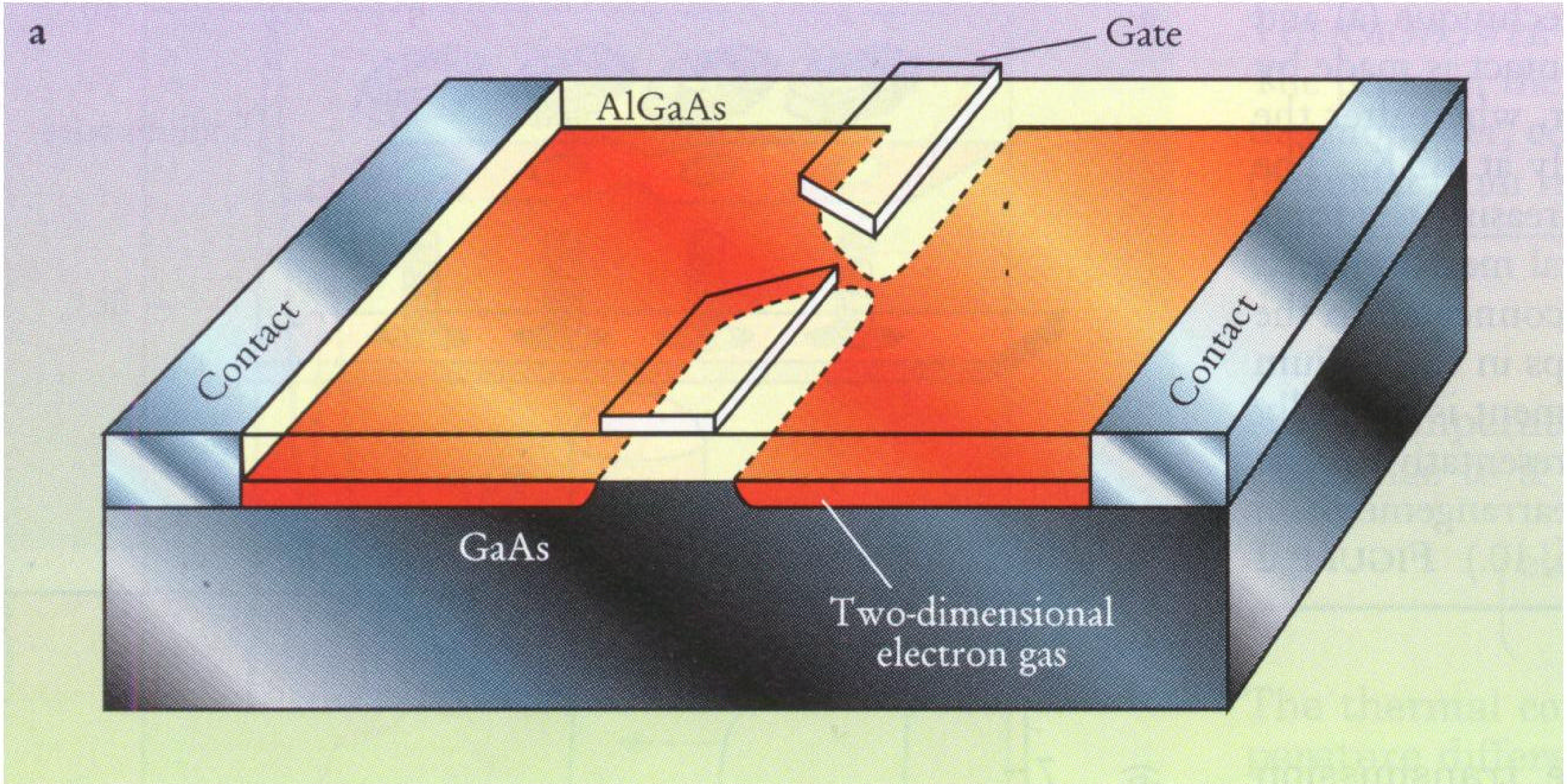
n = 1   n = 2



Free motion
(propagating waves)

Discrete states
(standing waves)

The quantum point contact: a solid state electron waveguide



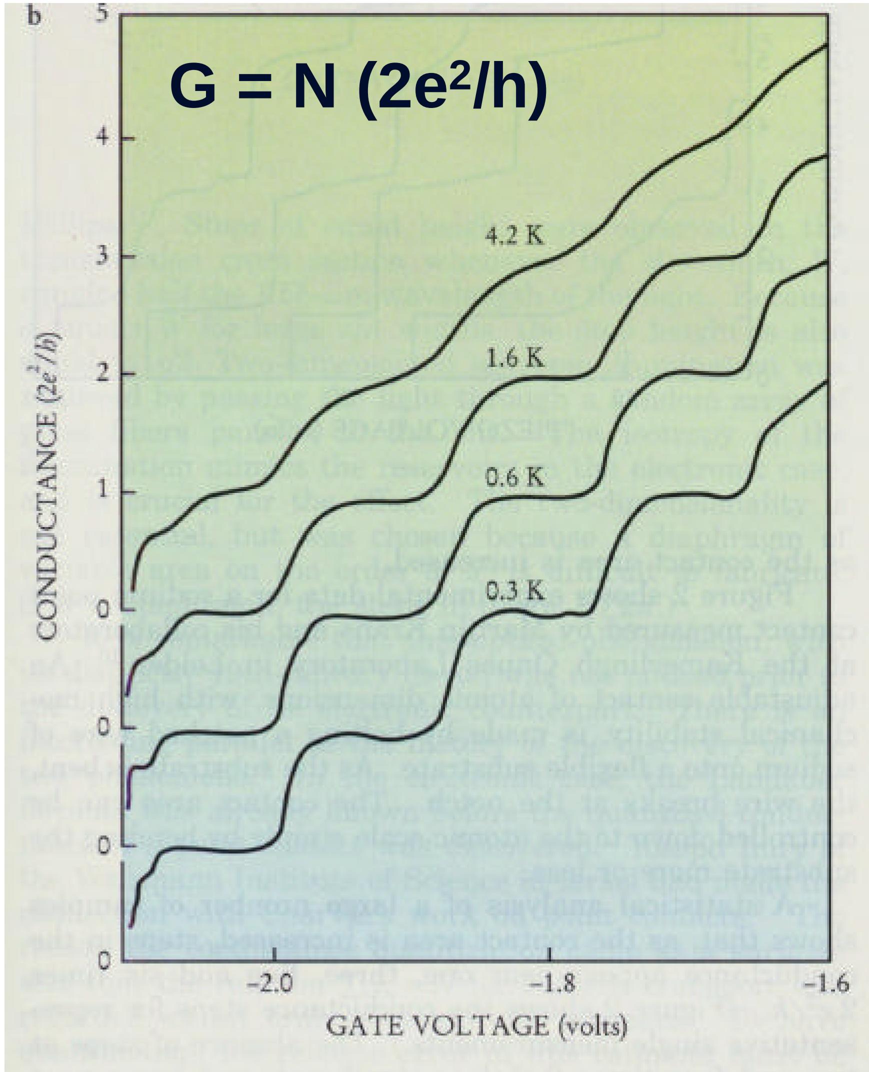
Henk van Houten and Carlo Beenakker, Physics Today, July 1996, p. 22-27

Conductance quantization

B.J. van Wees, H. van Houten, et al. Phys.Rev.Lett. 60, 848 (1988)

electron waveguide:
each occupied 1-d subband is a propagating mode and contributes ($2e^2/h$) to the conductance

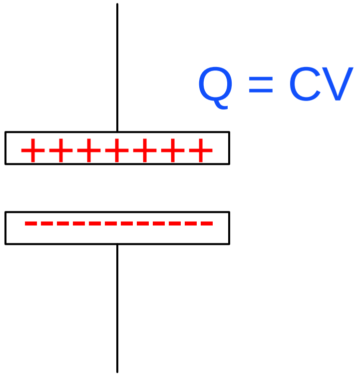
quantum ballistic transport:
visible if the 1d subbands are separated by more than kT



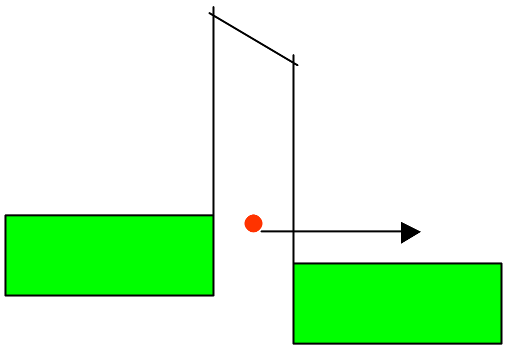
e^2 /h is a fundamental unit of conductance (cf. quantum Hall effect), it is the conductance of a single mode propagating from one **reservoir** to another

Single electron tunneling

Millikan 1911: charge is quantized, the elementary charge is *e*



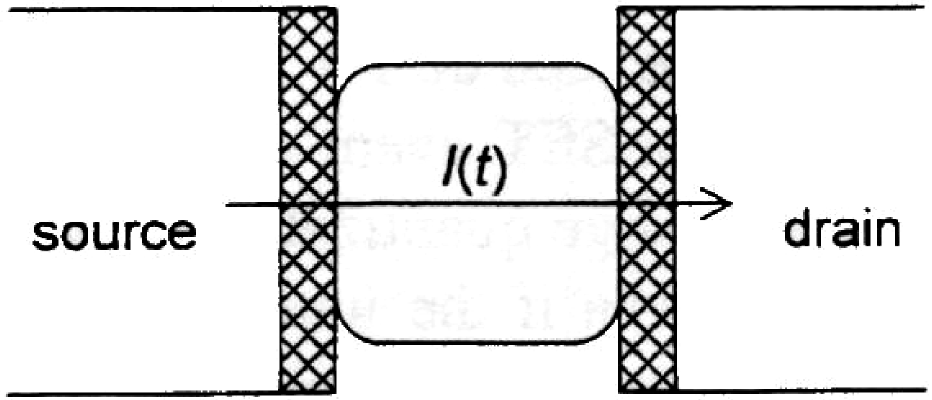
the charge induced on a capacitor can have any (fractional) value...



… but tunneling of electrons is a discrete process (manifested as shot noise with power 2e I)

Single electron tunneling

A metallic island is coupled by tunnel barriers to metallic leads.
The capacitance of the island with respect to the environment is C



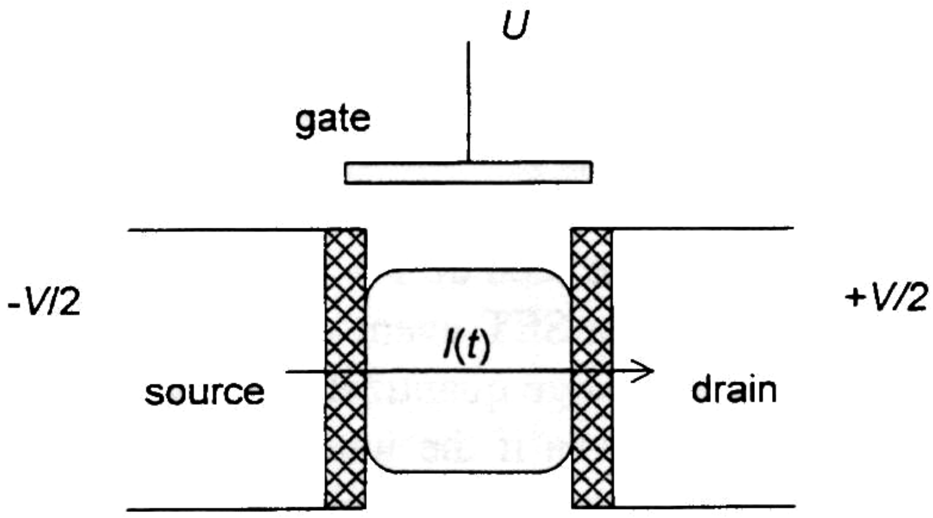
Number of electrons on an island is an integer if

- the elementary Coulomb charging energy $e^2/C >> kT$
- coupling to source and drain through tunnel barriers with resistance $R >> h/e^2$

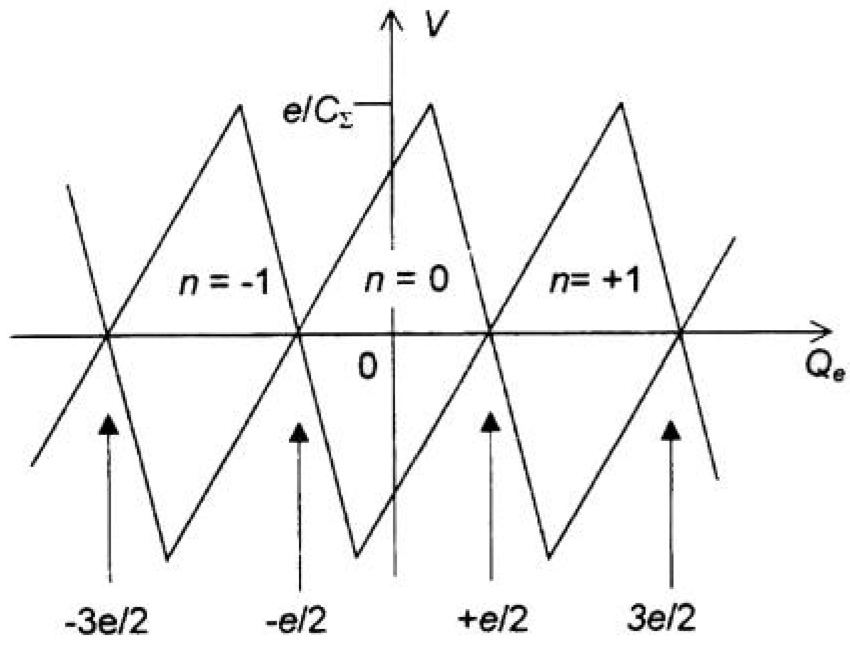
Coulomb blockade: no tunneling for small voltage!

K. Likharev, Proc. IEEE, 87 (1999) 606.

The single electron transistor (SET)



(a)



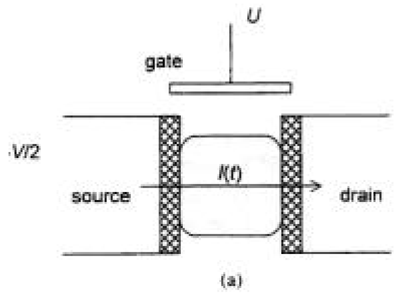
(c)

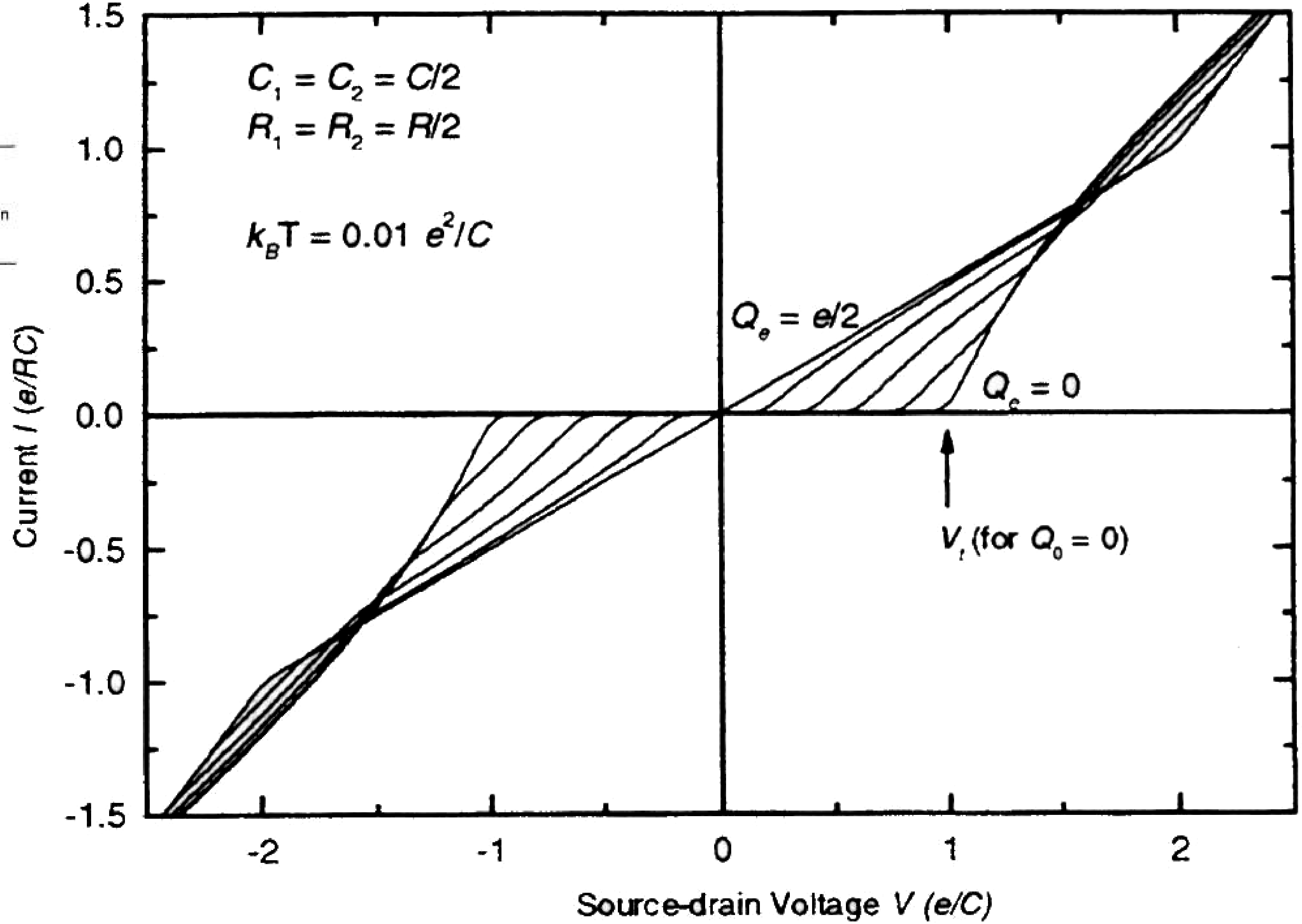
The threshold voltage V due to the Coulomb blockade oscillates as a function of the “induced charge” Q_e

K. Likharev, Proc. IEEE, 87 (1999) 606.

Henk van Houten, LATSIS symposium, June 2000

The single electron transistor (SET)

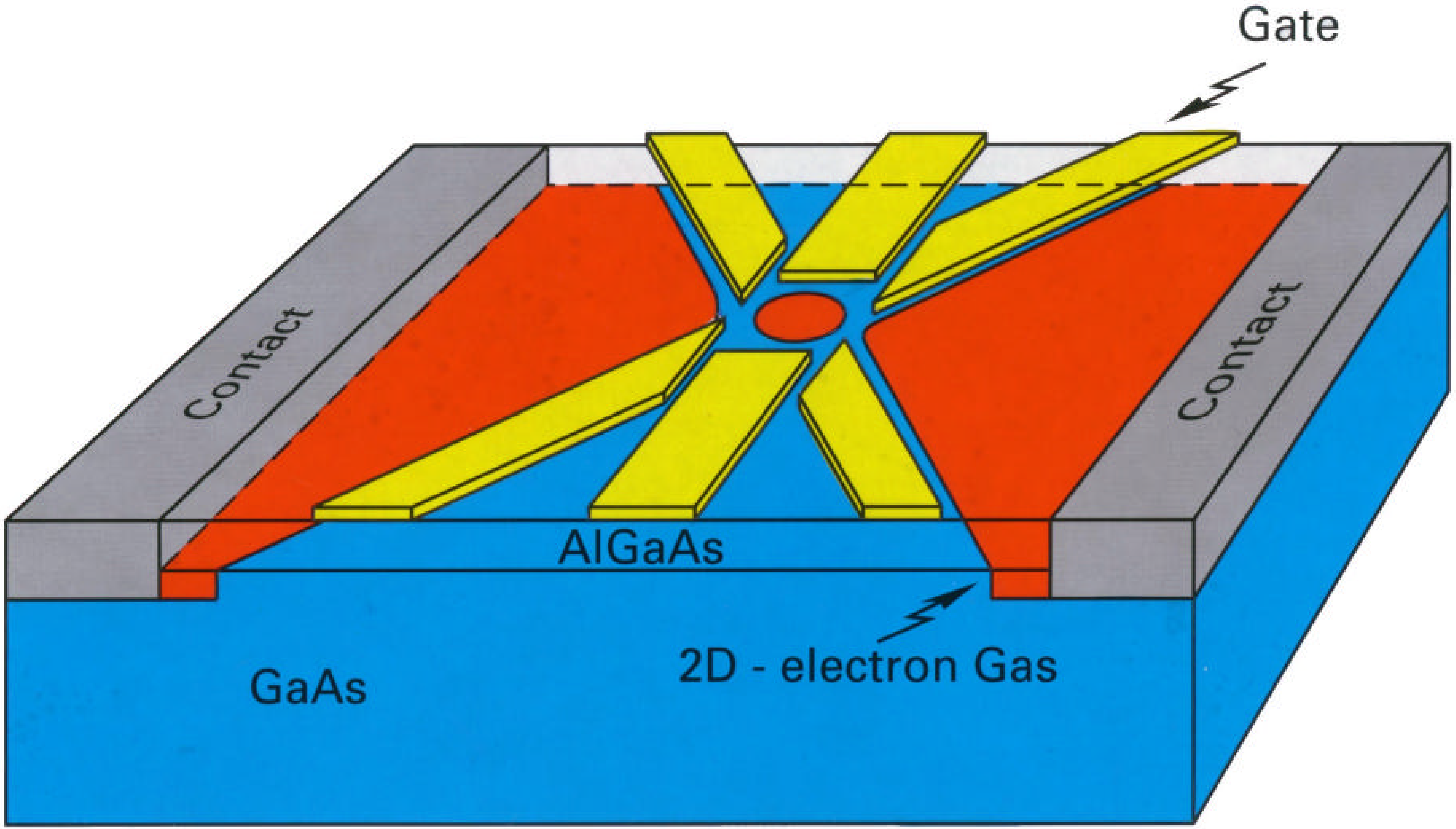




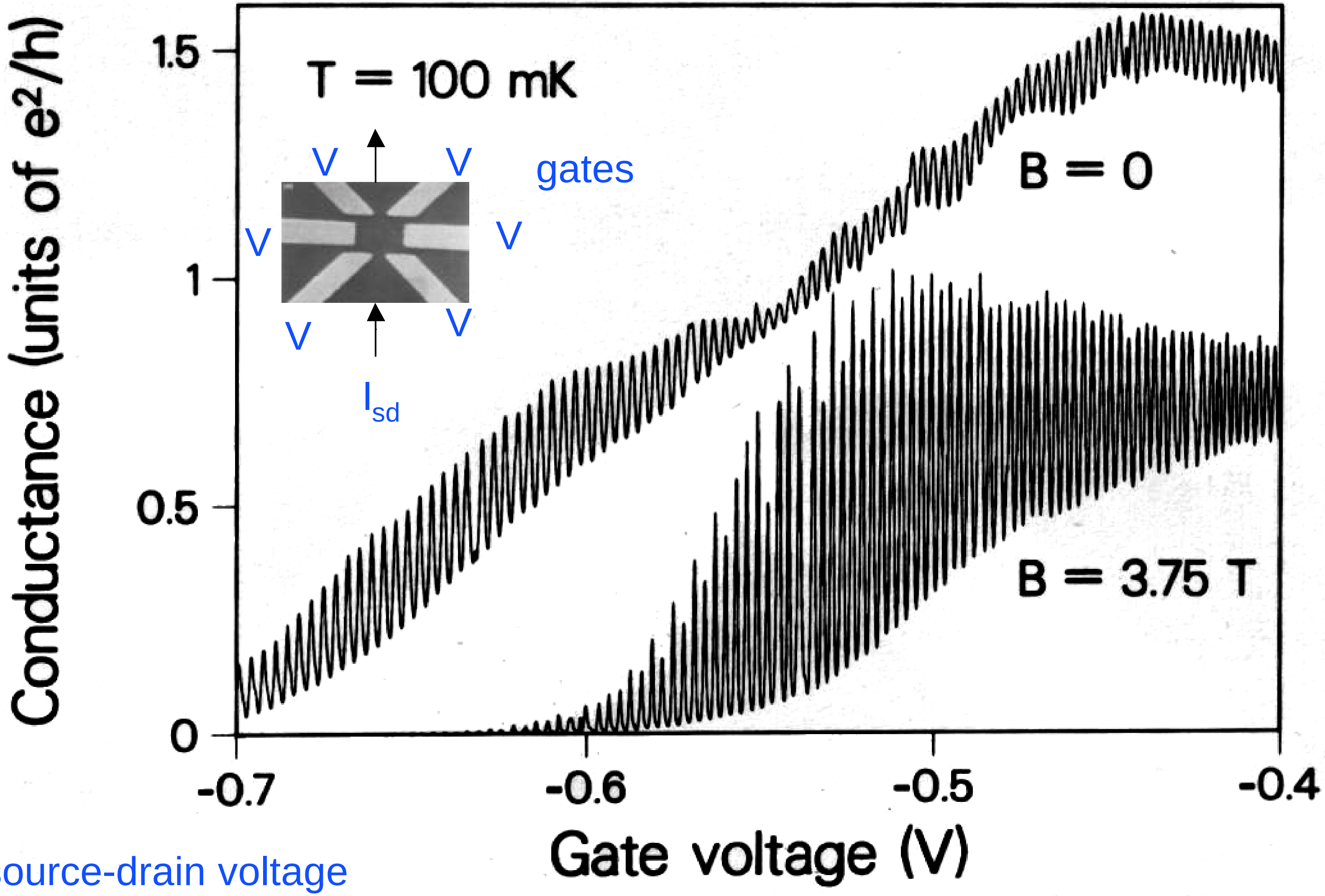
(b)

K. Likharev, Proc. IEEE, 87 (1999) 606.

The quantum dot



Coulomb blockade oscillations



Small source-drain voltage

H. van Houten, C.W.J. Beenakker, and A.A.M. Staring

Single Charge Tunneling, 1992.

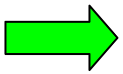
The ultimate MOSFET is a single electron device...

Minimum switching energy

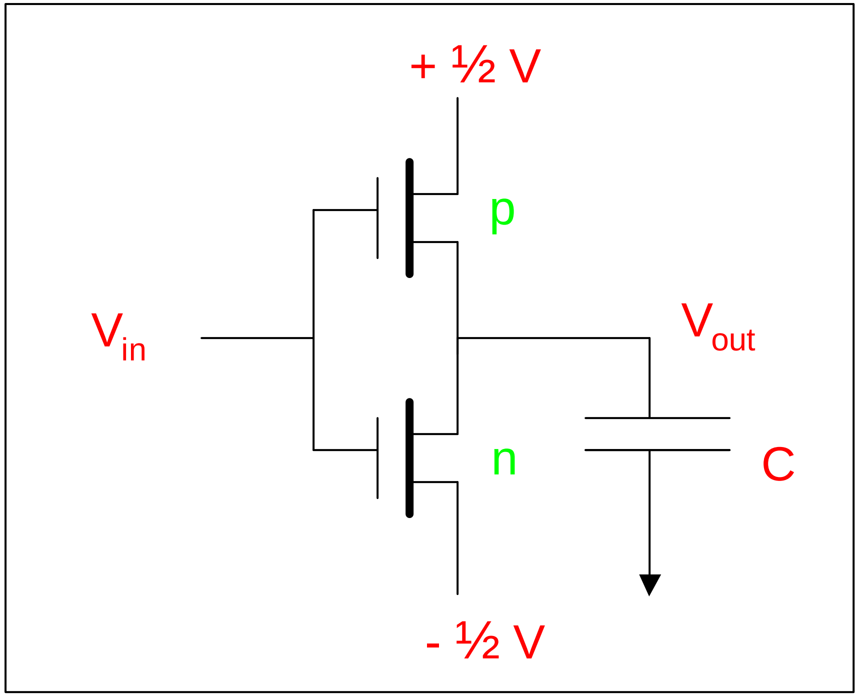
$$E_{switch} = \tfrac{1}{2}CV^2 > kT$$

Gain of inverter around threshold

$$V_{out}/V_{in} \sim eV / 2kT$$



$$V > kT/e\ , \quad E_{switch} \sim e^2/C$$



This corresponds to the *charging energy of a single electron*
The ultimate MOSFET thus is a “single electron MOSFET”

Example: L = W = 10 nm, oxide thickness 2 nm, in silicon.
It would switch in 0.1 ps (if velocity of the electron is the bulk Si saturation velocity)

J. Meindl, Proc. Int.Symp. Low Power Electronics and Design, Monterey, 1997, p. 149-151

Are quantum effects relevant for the ultimate MOSFET?

- MOSFET's are used far from equilibrium
- No quantum confinement between source and drain
- charging energy typically small compared to kT at room temperature (25 meV)

Quantum effects are "washed out"

There are some effects which modify the detailed behavior
hot electron effects
tunneling through the gate oxide
shifts in threshold voltage

The quantum limit of a switching device

The Heisenberg uncertainty relation

$$\Delta E \, \Delta t > h / 2\pi$$

imposes a quantum limit on the power-delay product for irreversible switching

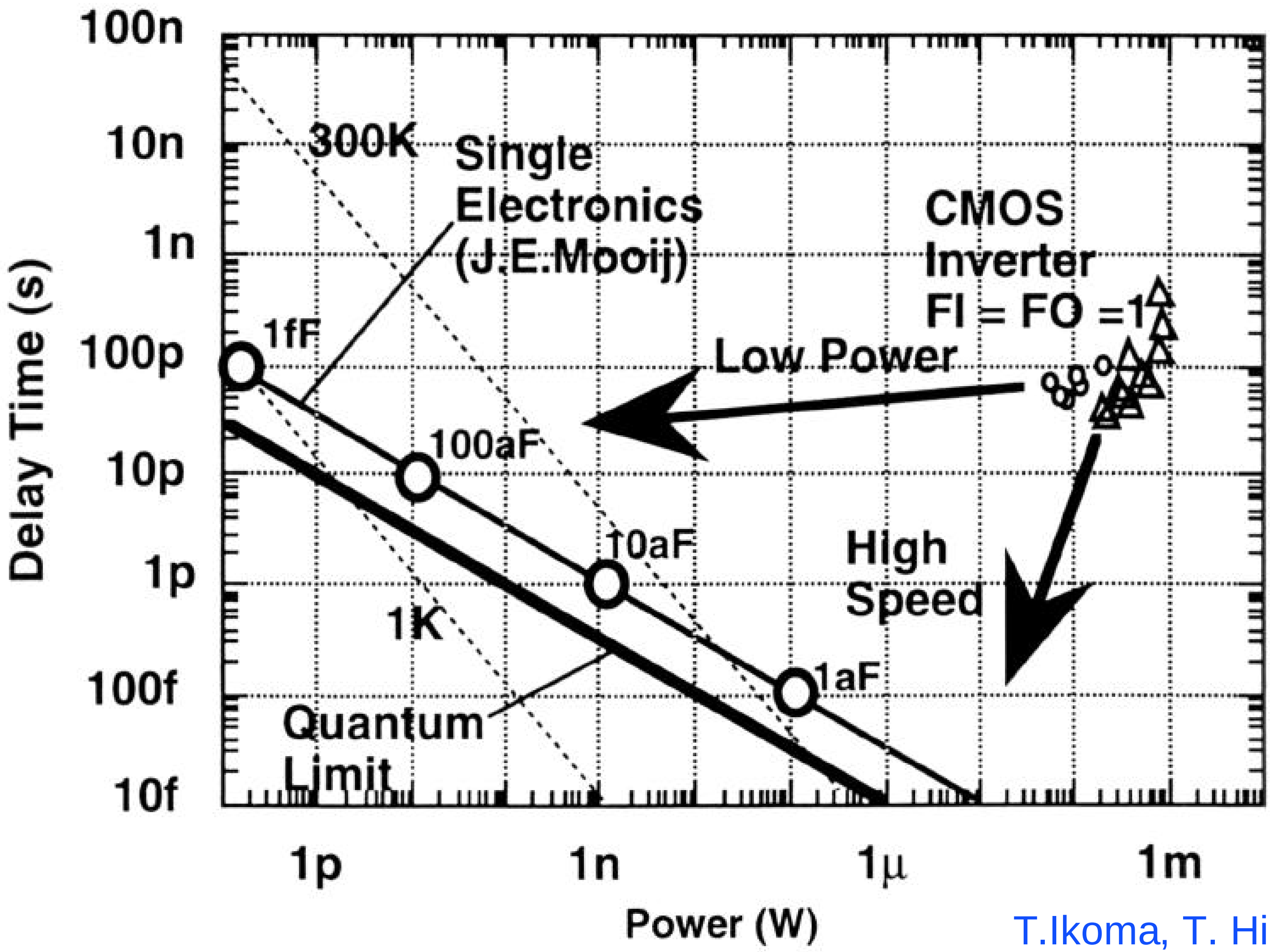
$$P \tau^2 > h / 2\pi$$

Single electronics is close to the quantum limit

$$\Delta E = e^2/C$$

$$\tau = RC \propto (h/e^2)C$$

Quantum limit and thermal limit



Quantum limit

$$P \tau^2 > h / 2\pi$$

Thermal limit

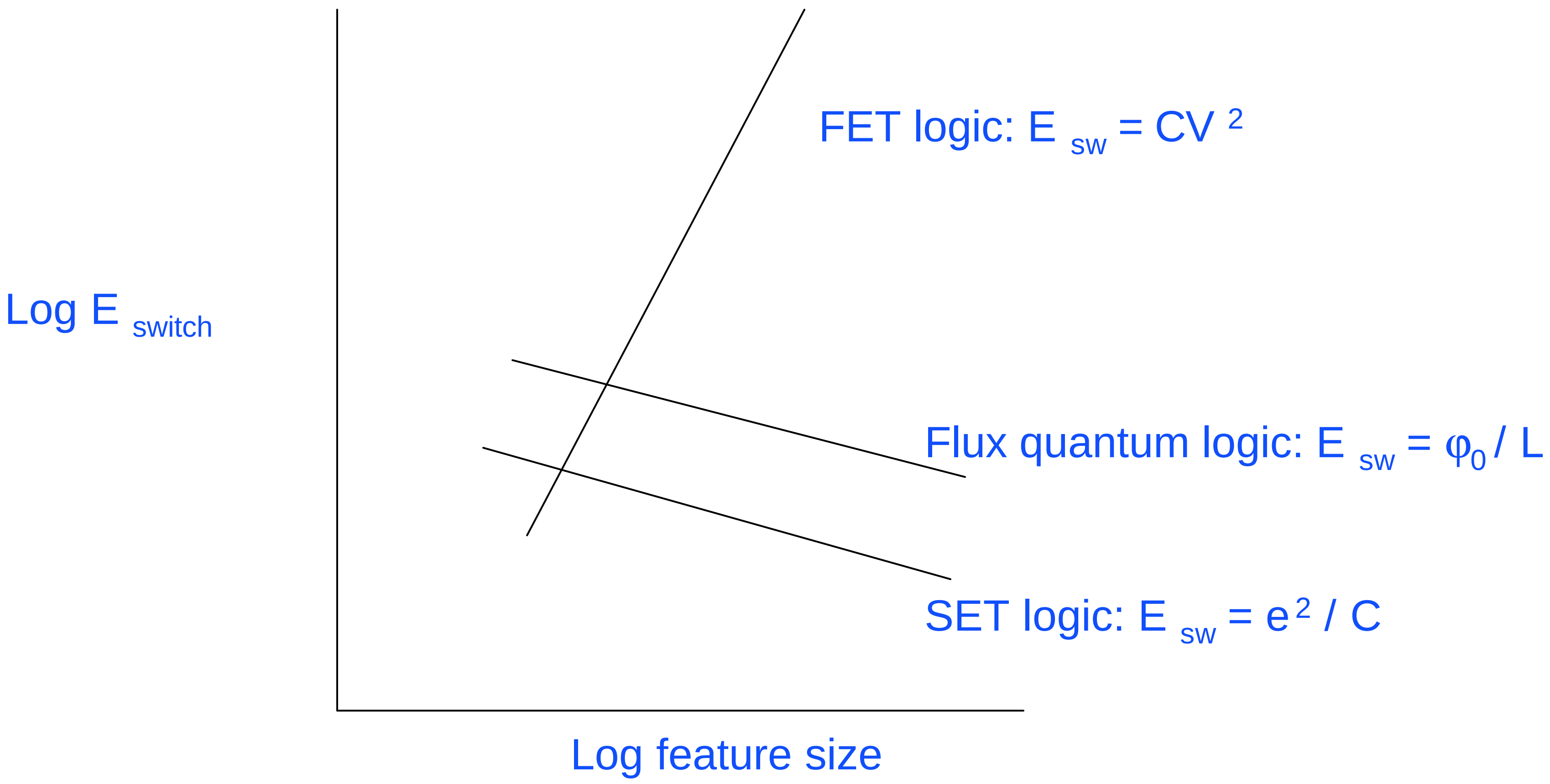
$$P \tau > kT$$

T.Ikoma, T. Hiramoto, K. Hirakawa

Can quantum effects be used for new types of devices?

- Replacement for MOSFET highly unlikely (see Likharev)
  - MOSFET still works fine, down to SE regime
  - SET devices have typically no gain, and no logic level restoration
  - offset charges lead to unpredictable offsets
  - making identical devices is nearly impossible
  - multi-valued response is undesired for conventional architectures
  - $h/e^2 \approx 25\ k\ \Omega$, poor matching to impedance of transmission line, and leading to large RC time for charging the interconnect

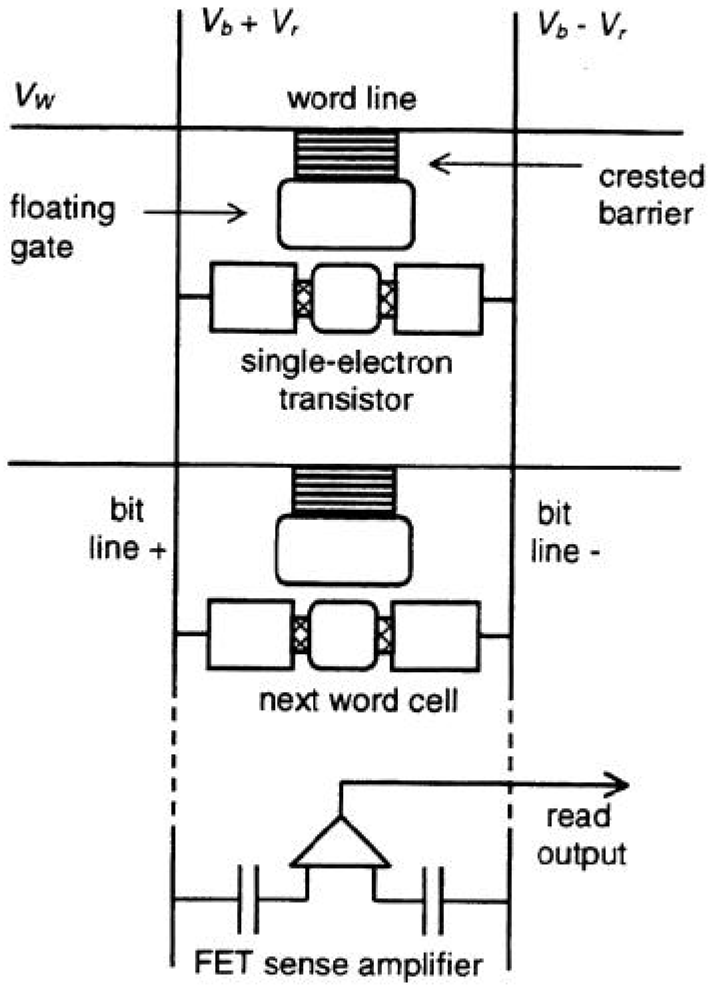
Scaling for quantum devices



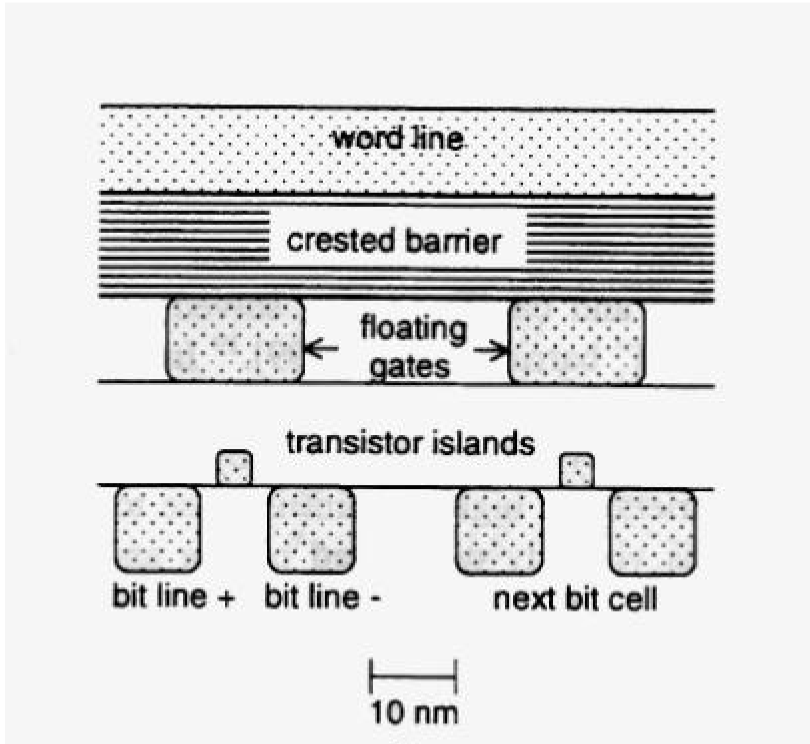
Quantum devices dissipate more if you scale them down

(Carver Mead and Lynn Conway, VLSI Systems)

SET-FET hybrid memory cell



(a)



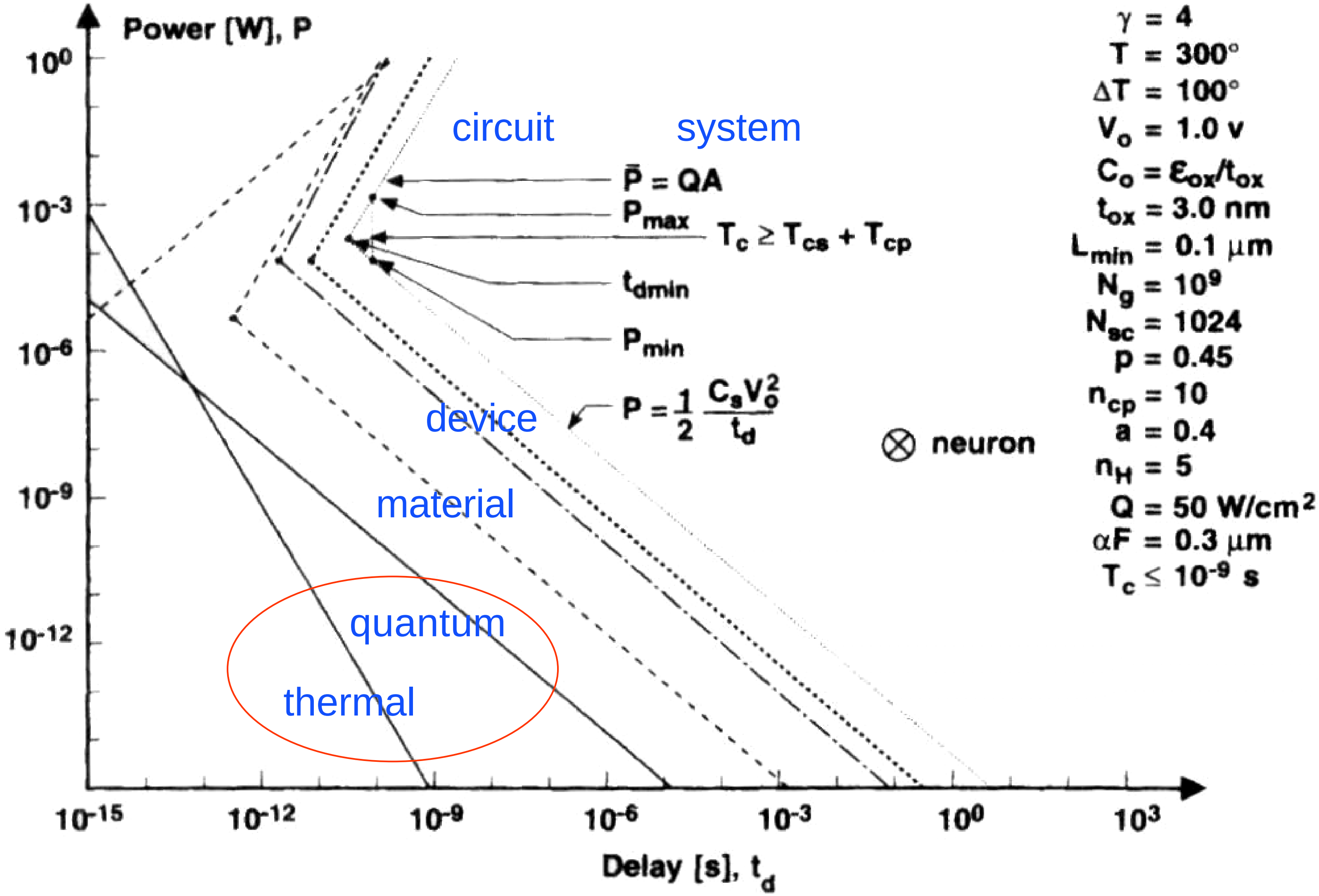
(b)

K. Likharev, Proc. IEEE, 87 (1999) 606.

Contents

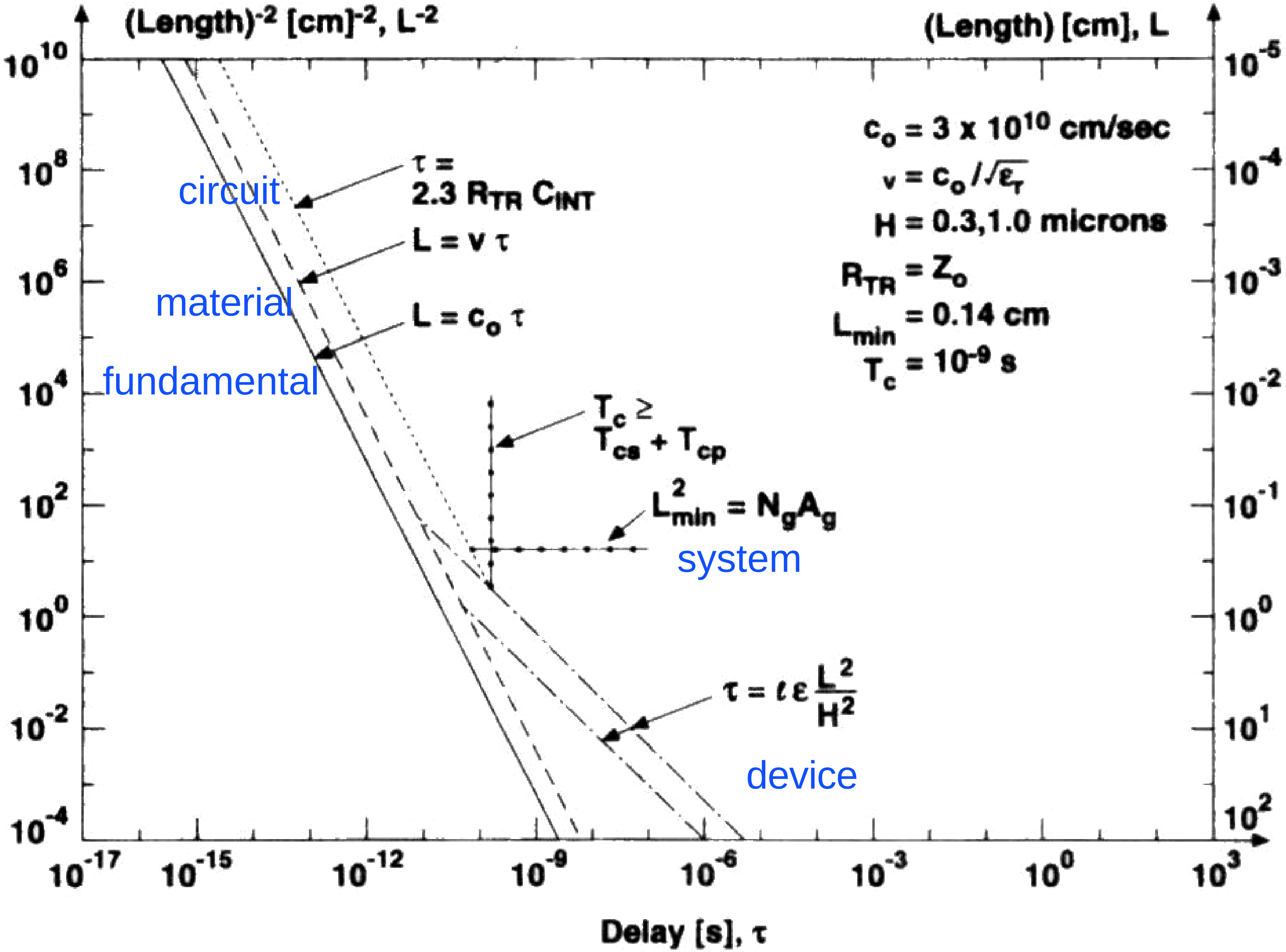
- Digital switching and the thermal limit
- Digital switching and the quantum limit
- **A hierarchy of limits (Meindl’s classification)**

Meindl on switching limits



Henk van Houten, LATSIS symposium, June 2000

Meindl on interconnect limits (case study)

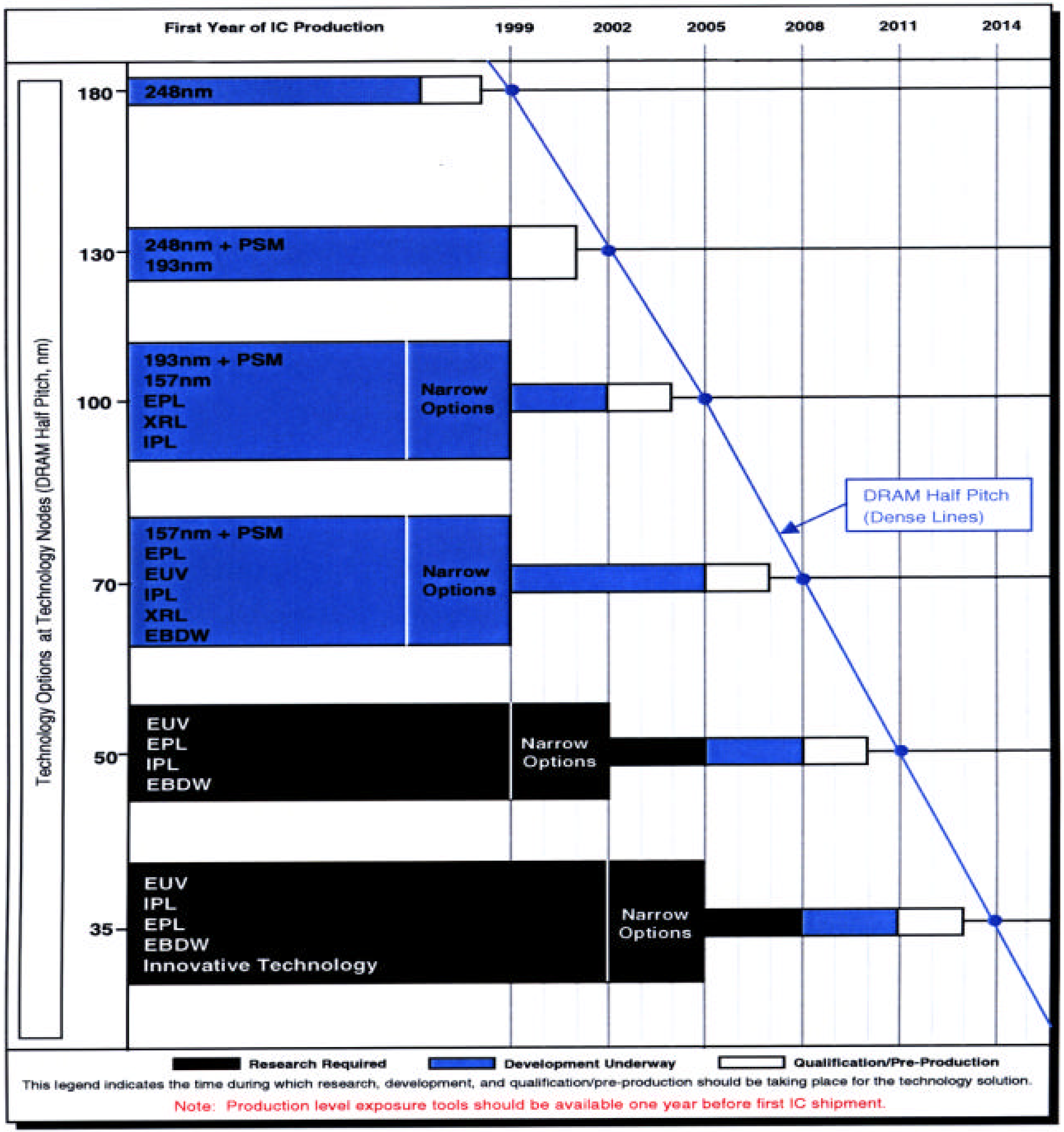


Meindl's "hierarchy of limits"

level	*limits*
System	Ultimate system (?) 1 billion gates, 0.1 mu CMOS, Q = 50 W/cm^2, clock 1 ns
Circuit	Transfer curve, switching energy, propagation delay, global interconnect response time
Device	Ultimate MOSFET? L =50 nm, tox=3 nm E = 0.014 fJ = 87 eV T < 0.5 ps
Material	Saturation velocity Dielectric constant Breakdown field Thermal conductivity
Fundamental	Thermodynamics Quantum mechanics Electromagnetism

Lithography (Sematech roadmap)

Two major contenders:
EUV(13 nm) and e-projection

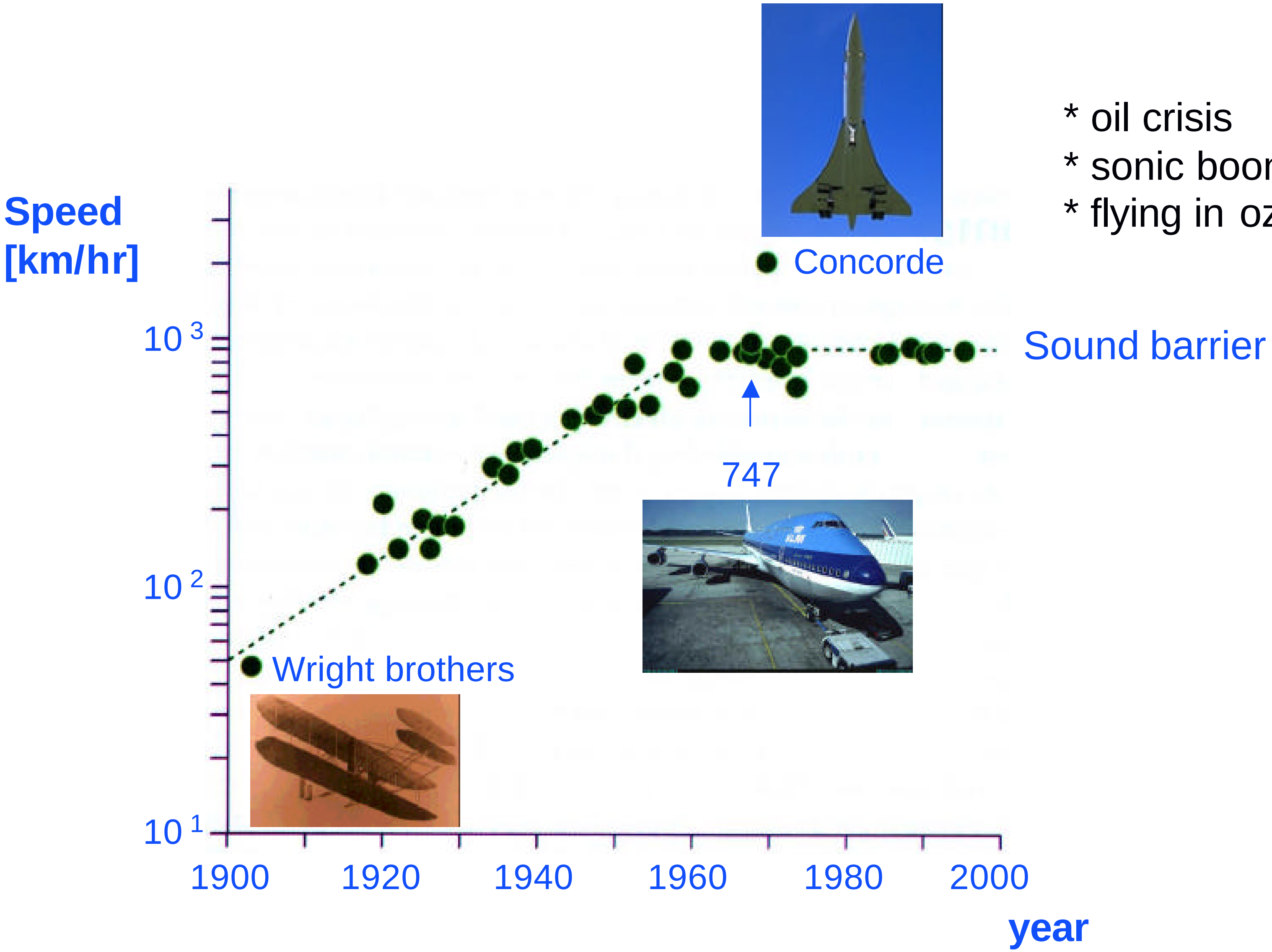


Henk van Houten, LATSIS symposium, June 2000

Why Moore's law may break down (say in 2014, @ 1 Tbit DRAM)

- lithography
  - 35 nm node, 2 nm CD control for a MPU, 15 nm overlay, mask making tremendously difficult, mask and tool cost
- process technology and yield
  - gate oxide thickness <1 nm, fluctuations in doping profiles (100 atoms long gate length, 100 dopant atoms)
- power dissipation
  - high performance: heating of the chip
  - portable: battery life
- (global) interconnects
  - increasing propagation delay & parasitics
- design complexity
- economical factors

Historic trend of aircraft speed



* oil crisis
* sonic boom
* flying in ozone layer (NOx)

Source: Nederlands Tijdschrift voor Natuurkunde 1997

Conclusions

- Information/computers have a physical basis
  - scaling of FET transistors is at the basis of the IT revolution
- Common wisdom physical limits are not really fundamental ...
  - Feynman 1985: “these are the only physical limitations on computers that I know of”
    - limitations to the size of atoms
    - energy requirements depending on time
    - speed of light
- ...but quantum devices seem to offer mainly disadvantages
- Practical limits and economical considerations are likely to determine how far we can stretch Moore’s law (2014?)

Key references

- Anthony J.G. Hey, ed. *Feynman and Ccmputation* (Perseus, Reading MA, 1999)
- Carver Mead, Lynn Conway, *Introduction to VLSI systems*, Addison Wesley, ...
- Rolf Landauer, *Dissipation and noise immunity in computation and communication*, Nature, 335, 779-784, (1988)
- James D. Meindl, *Low power microelectronics: retrospect and prospect*, Proc. IEEE, 83, 619-635 (1995); *Interconnect limits on XXI century Gigascale Integration*, Mat. Res.Soc. Symp. Proc. Vol. 514, 3-9, 1998
- Paul M. Solomon, *Critique of reversible computing and other energy saving techniques*, in *Future Trends in Microelectronics-Reflections on the Road to Nanotechnology*, ed. Serge Luryi, Jimmy Xu, and Alex Zaslavsky, NATO ASI, E 323, p. 93-109
- Carlo W.J. Beenakker and Henk van Houten, *Quantum Transport in Semiconductor Nanostructures*, Solid State Physics, 44 (1991), p. 1-228.
- Henk van Houten, Carlo W.J. Beenakker, and A.A.M. Staring, in Hermann Grabert and Michel H. Devoret, eds, *Single Charge Tunneling-Coulomb Blockade Phenomena in Nanostructures*, NATO ASI, B 294. (Plenum, New York, 1992)
- Konstantin K. Likharev, *Single electron devices and their application*, Proc. IEEE, 87, 606-632 (1999).
- Charles H. Bennett and David P. DiVincenzo, *Quantum Information and computation,* Nature, 404, 247-255, 2000