Galileo, Elephants, & Fast Nano-Devices

Mark Rodwell
University of California, Santa Barbara

rodwell@ece.ucsb.edu 805-893-3244, 805-893-5705 fax

Scaling: making transistors small makes them fast

We've recently made very fast transistors...
...mostly by making them small.

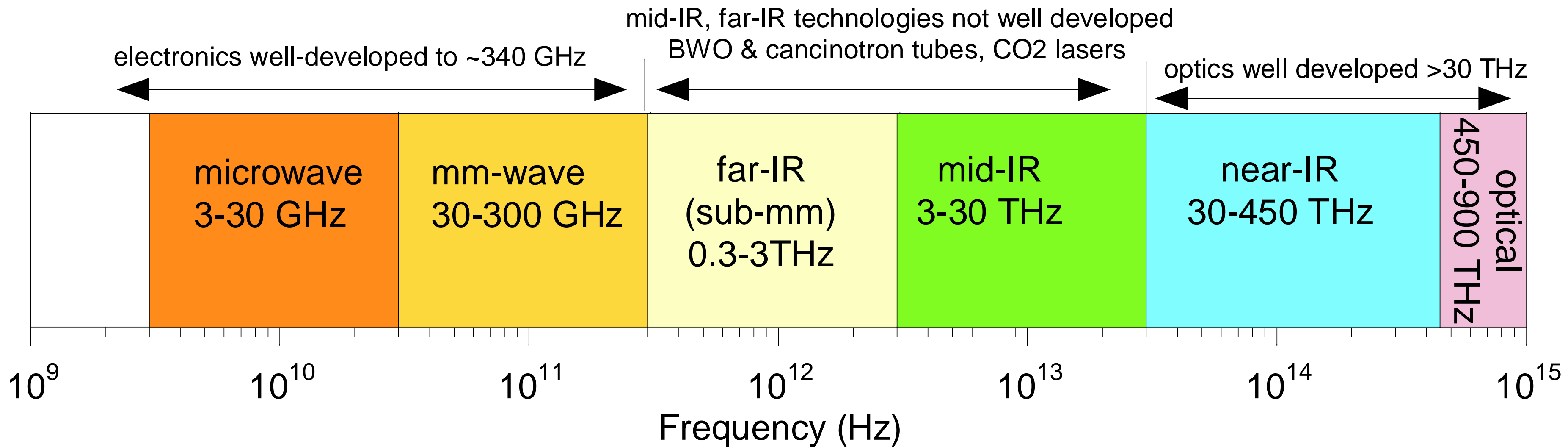
This is related to Galileo and to elephants

So:

what are transistors ?
what are they for ?
how do they work ?
what limits their speed ?
why does making them small help ?

...and how high in frequency can electronics work ?

Goal: Make ICs which work in the Infrared



Far-IR and Mid-IR sources / detectors today:
BWO & carcinotron vaccum tubes, CO_2 & quantum cascade lasers

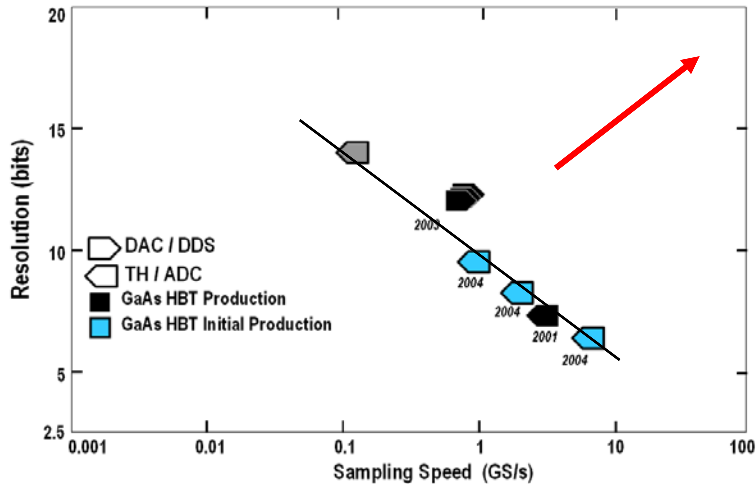
But: while these do <u>make power</u> at THz frequencies,
they can't <u>process signals</u> at comparable rapidity, and <u>don't do much else</u>

Our goal: <u>Transistors</u> and <u>Integrated Circuits</u> for 300-1000 GHz
tiny very sensitive (low noise) very rapid modulation (many bits/second)

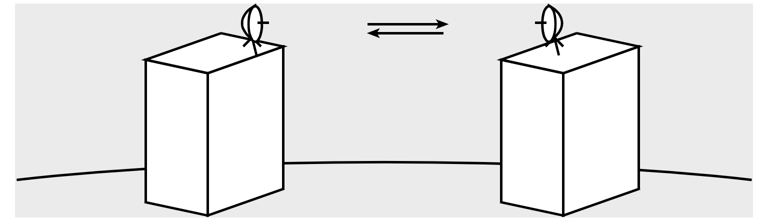
Transistors → ICs → very complex signal processing being done very very quickly.

What could we do with a 5 THz Transistor ?

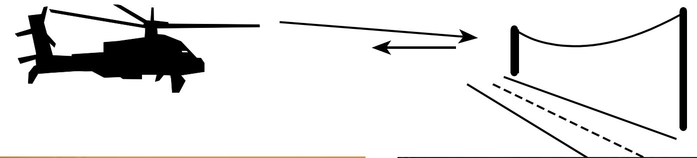
High-Resolution Microwave ADCs and DACs



sub-mm-wave radio:

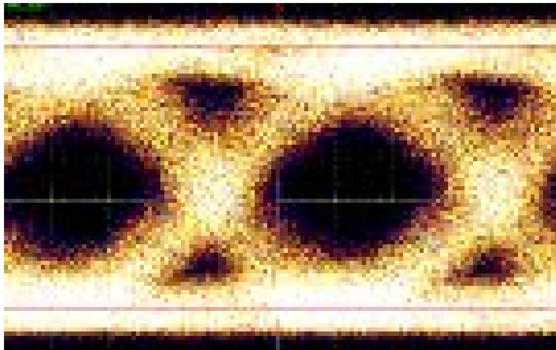


340 GHz & 600 GHz imaging systems

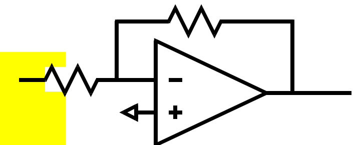


320 Gb/s fiber optics

& adaptive equalizers for 40 Gb/s ...

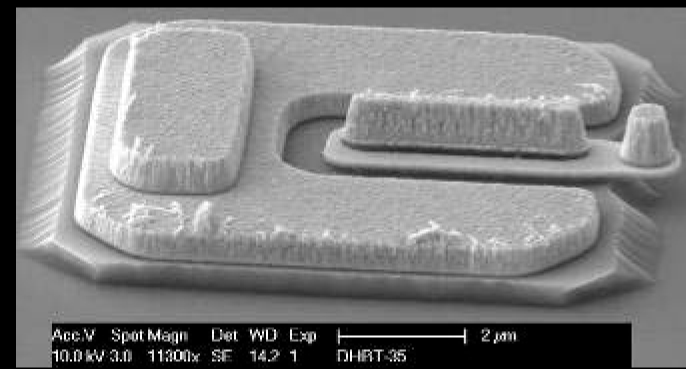


Precision Analog design
at microwave frequencies

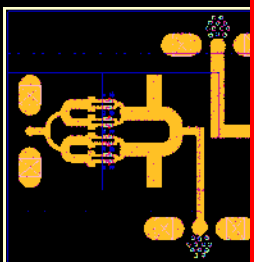


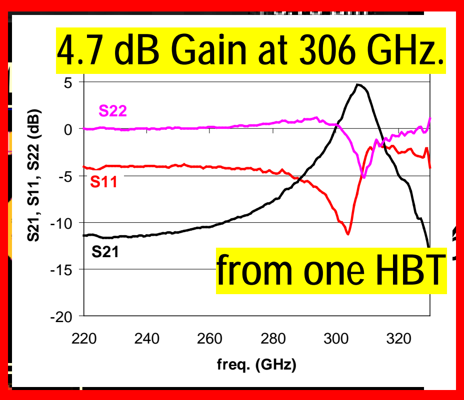
Why develop THz transistors ?
→ compact ICs supporting complex high-frequency systems.

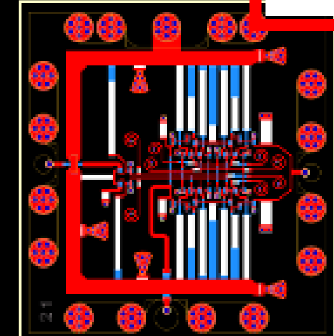
Tiny Transistors Are Very Fast Transistors



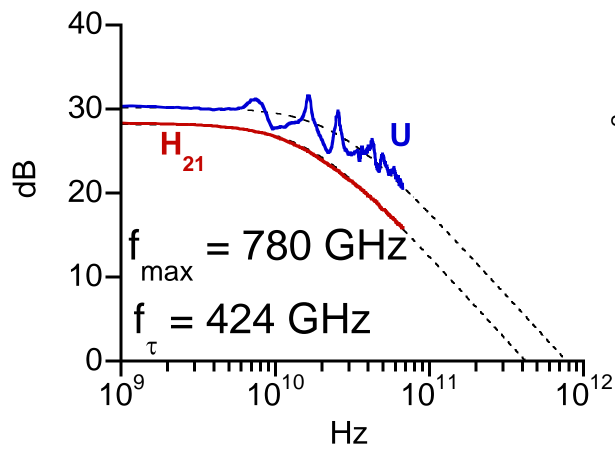
340 GHz, 70 m

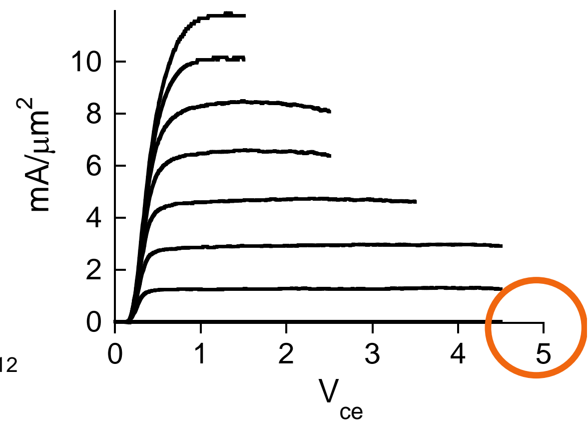


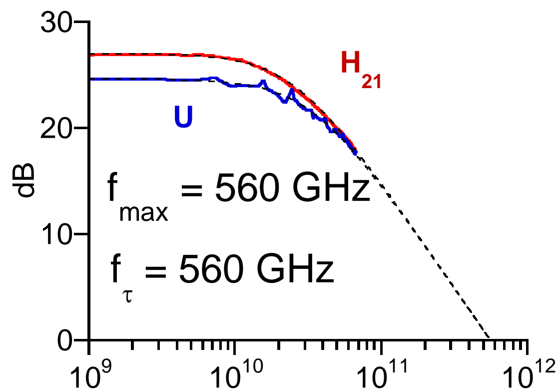


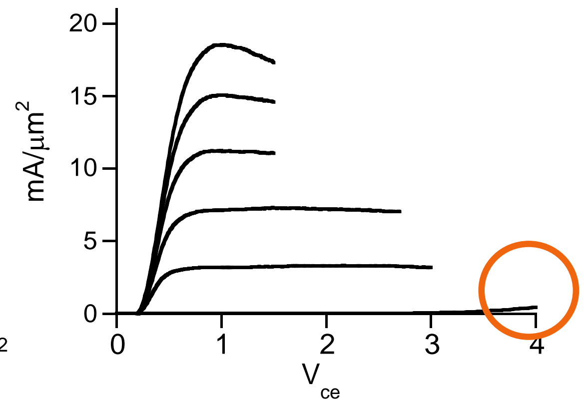


200 GHz master-slave latch design

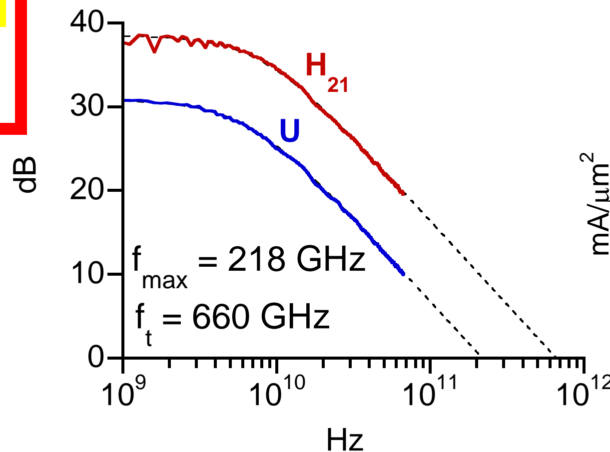


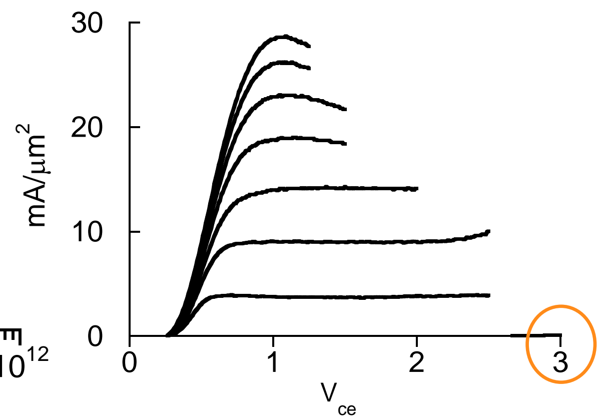






0 nm thick collector





Z. Griffith, E. Lind, J. Hacker, M. Jones

First Consider Scaling... & Elephants

10:1 (taller /wider/ deeper)

1000: 1 more metabolism, 100:1 larger skin area surface→ overheats

1000: 1 larger weight, 100:1 larger bone cross-section→ legs break

1000: 1 more flesh, 100:1 larger lung surface→ suffocates

(plagiarized from Galileo)

Scaling... a golf ball

$\left(\dfrac{\text{volume}}{\text{surface area}}\right)$ *ratio has changed a bit*

Scaling: *little things change more quickly than big things*

Scaling:

the surface matters most in little things,
the bulk matters most in big things

Ground Rules

"Everything should be made as simple as possible, but not simpler."

(Einstein)

We can simplify, but not to the point where we ignore key considerations.

Enthusiasm enables, hype mis-directs...

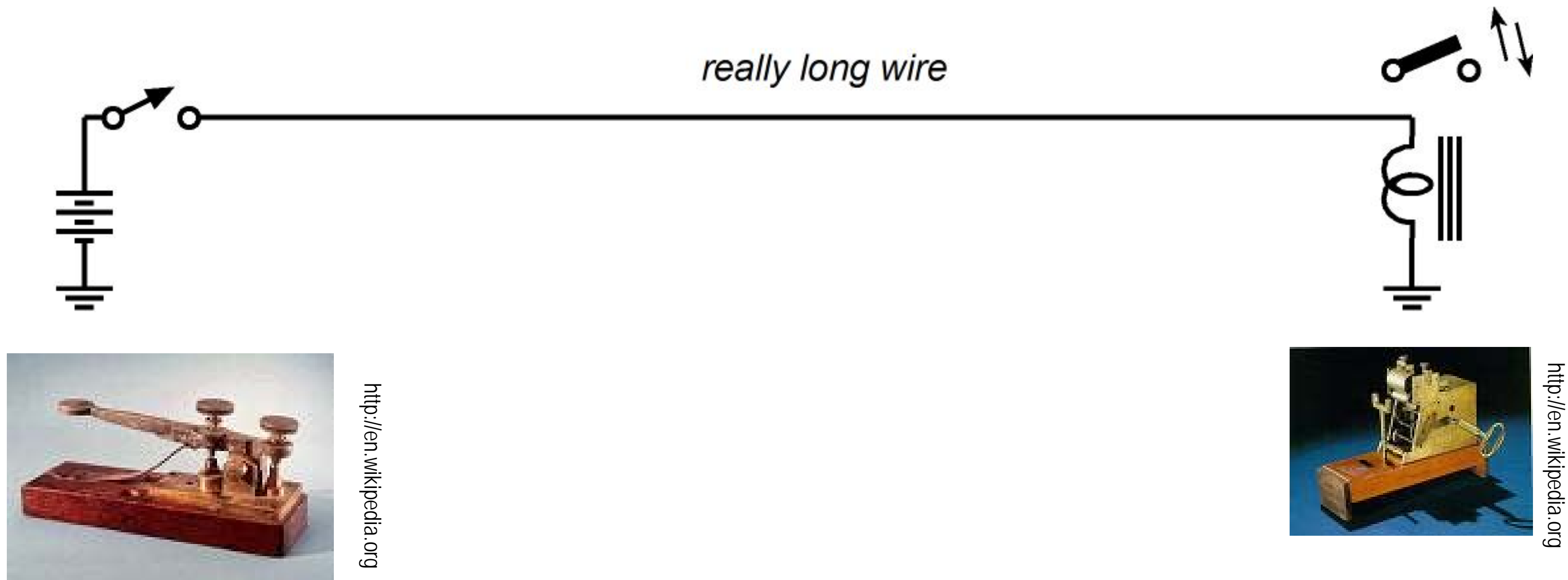
Tubes & Transistors

...what are they ?

...what are they for ?

...how do they work ?

The Telegraph: The First Electronics (1830's)



Schilling , Morse, Wheatstone, Edison, Gauss, Heaviside...

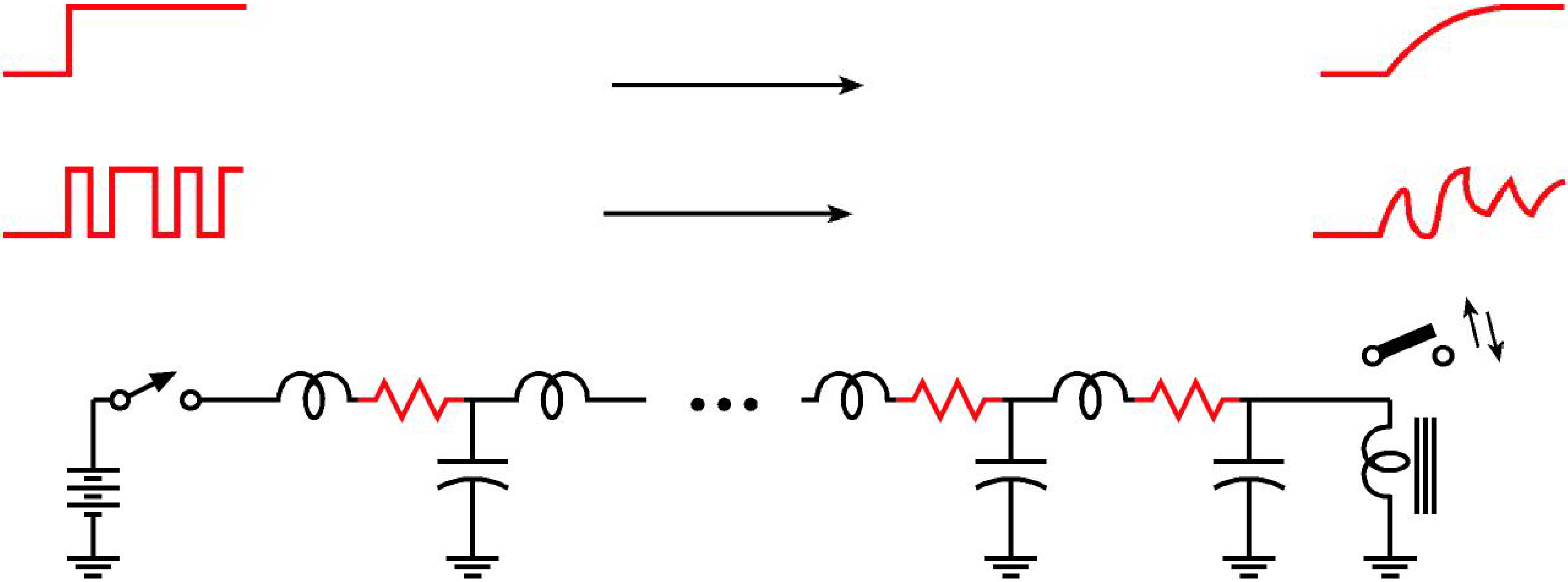
"The Ancients have Stolen Our Inventions"

pulse dispersion, frequency-division multiplexing

Frequency-domain transform methods

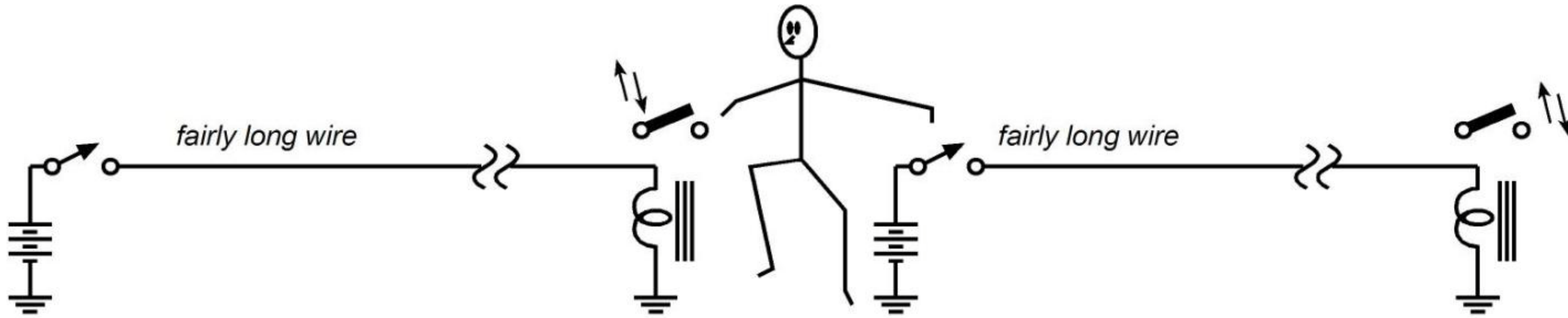
amplification

Loss and Dispersion Limits Range



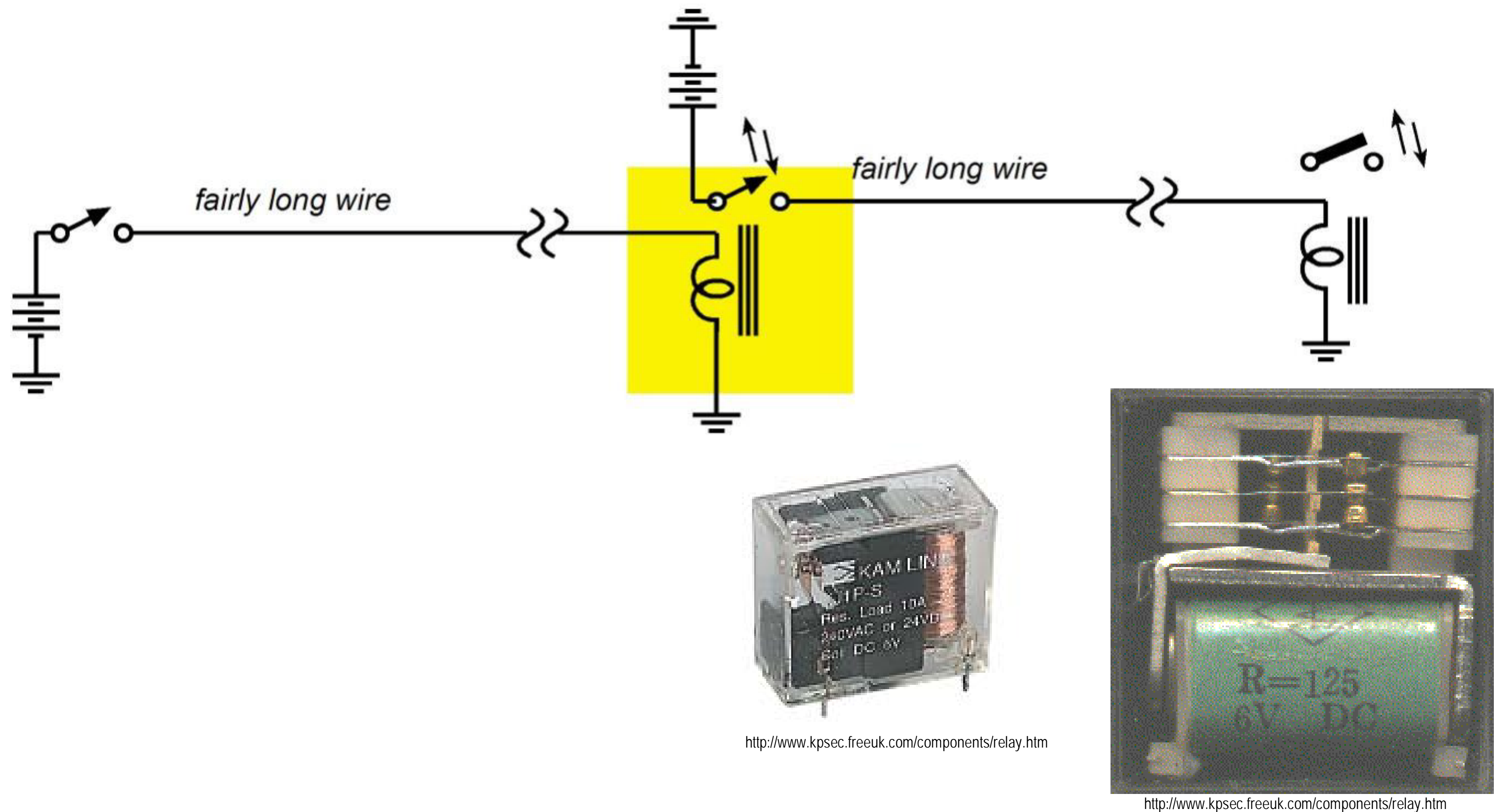
Resistance→ **pulse dispersion**
The longer the range, the more slowly you must signal

Human Relay To Repeat the Signal



Expensive and Slow…

Magnetic Relay: the First Electrical Amplifier



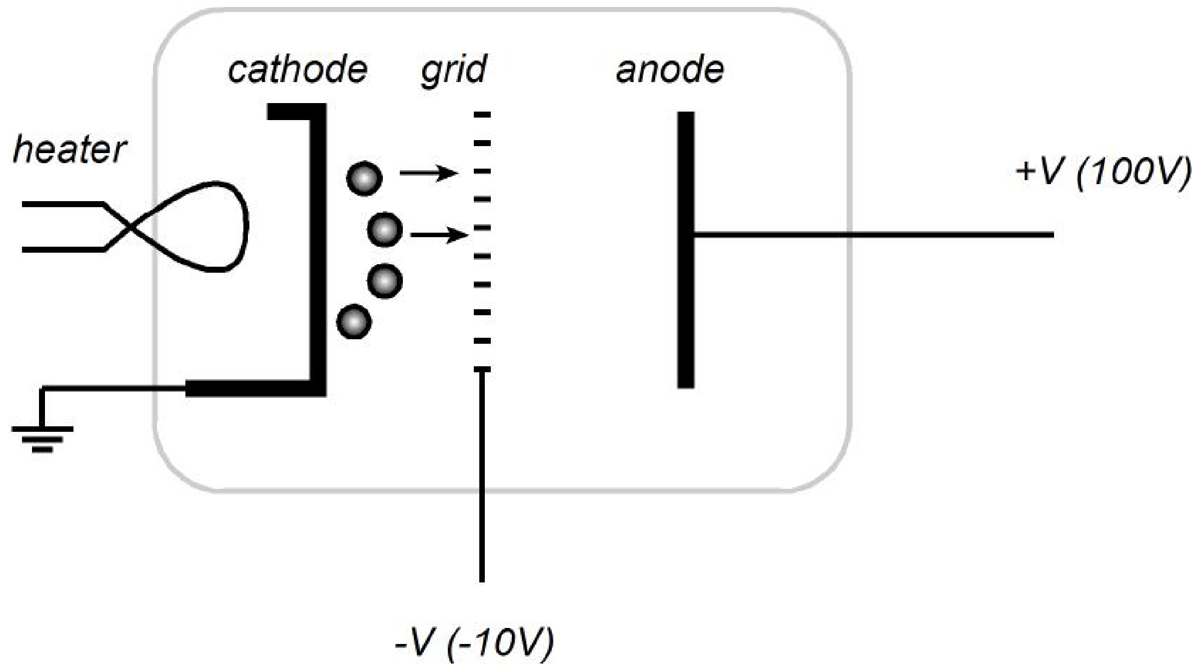
http://www.kpsec.freeuk.com/components/relay.htm

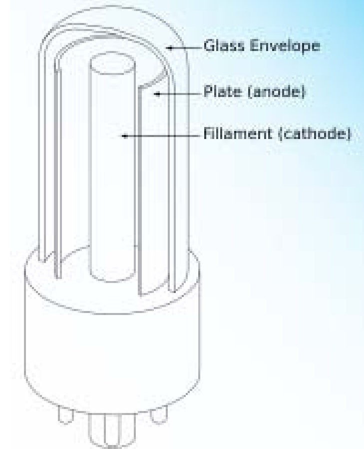
http://www.kpsec.freeuk.com/components/relay.htm

Question asked when "Tubes" or "Valves" were first introduced:
"Is it a true relay ?" ---- meaning: "Is it an amplifier ?"

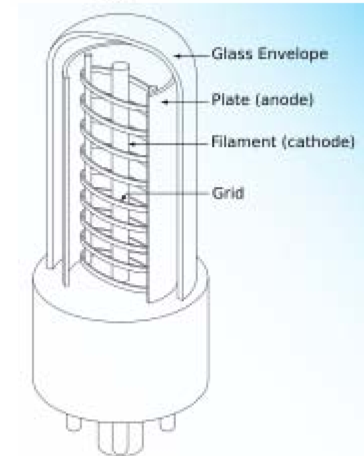
Modern terminology:
"Is there {voltage, current, power} amplification?"

Vacuum Tubes (1903-1907) Edison, Thompson, Fleming, DeForrest





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



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

How it works:

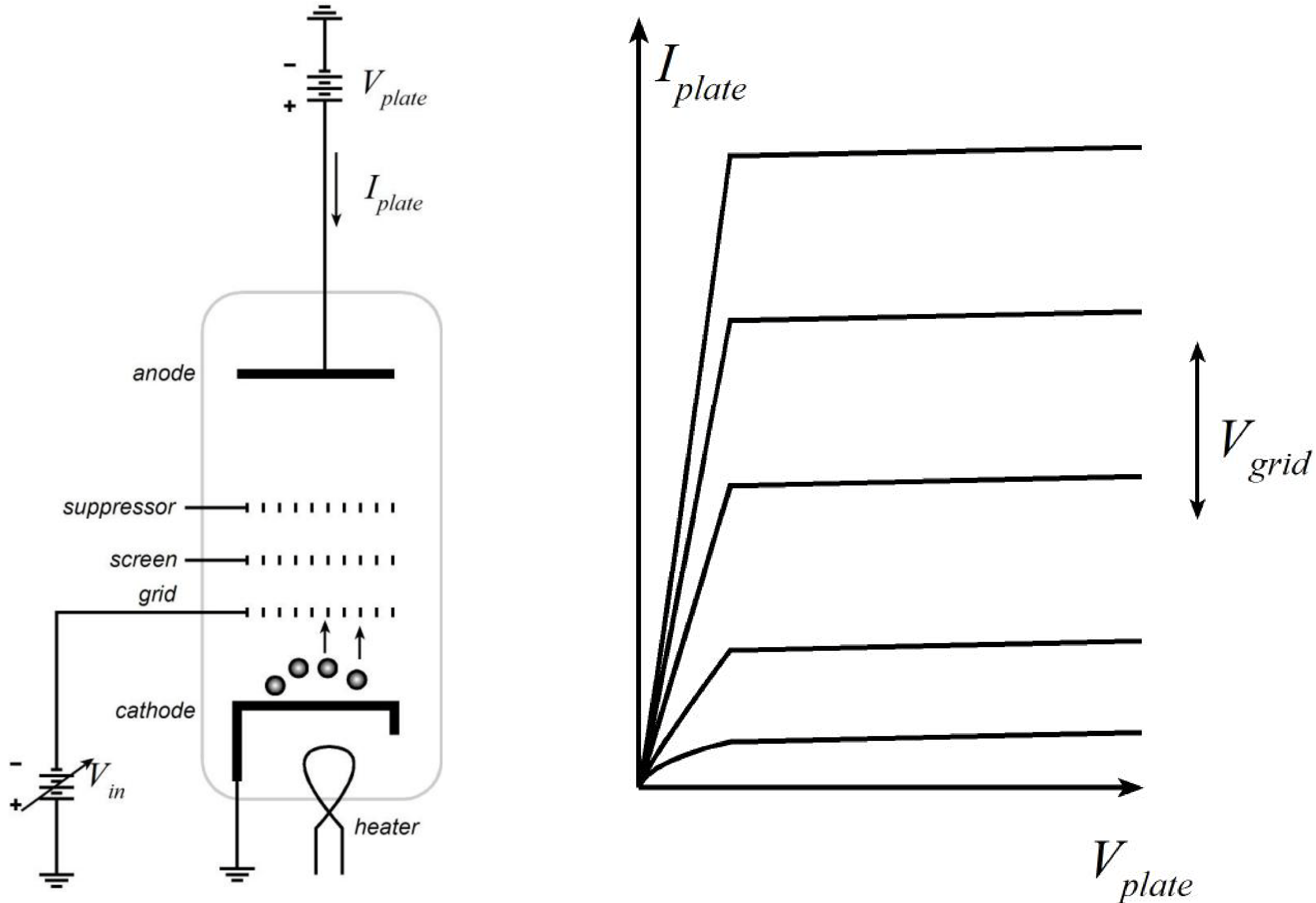
Hot cathode boils electrons into Vacuum

Grid screens electrons near cathode from positive anode

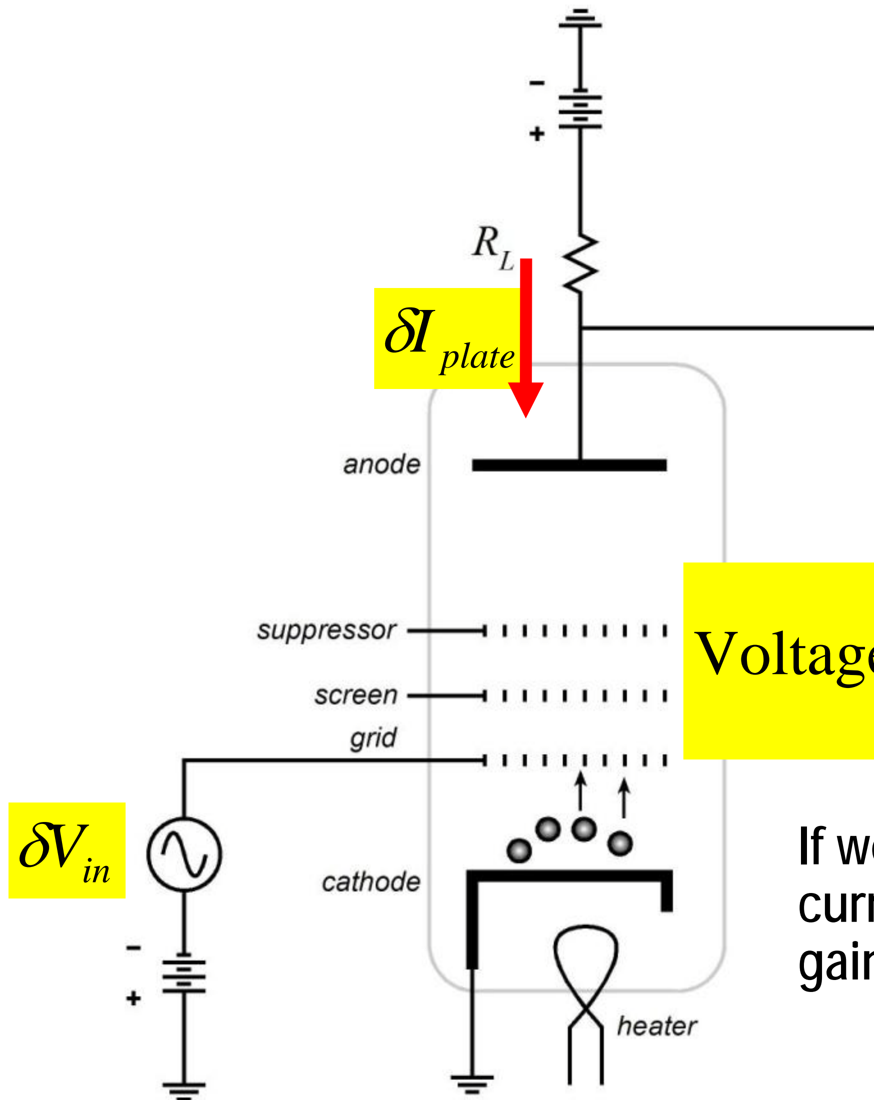
Negative grid repels electrons: the more negative, the less current

Electrons passing through grid drawn quickly to Anode

Tubes: Input Voltage Controls Output Current



Vacuum Tubes --- As an Amplifier

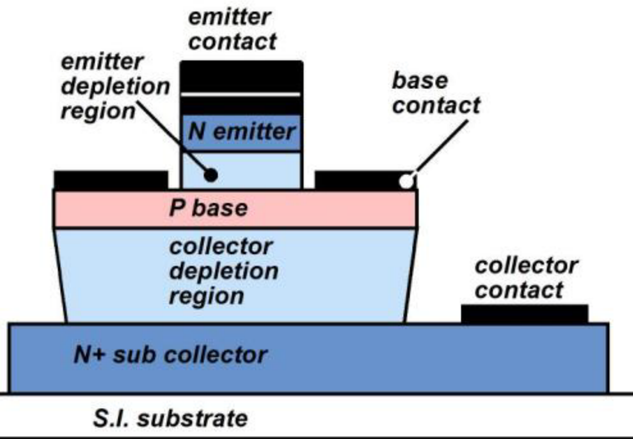


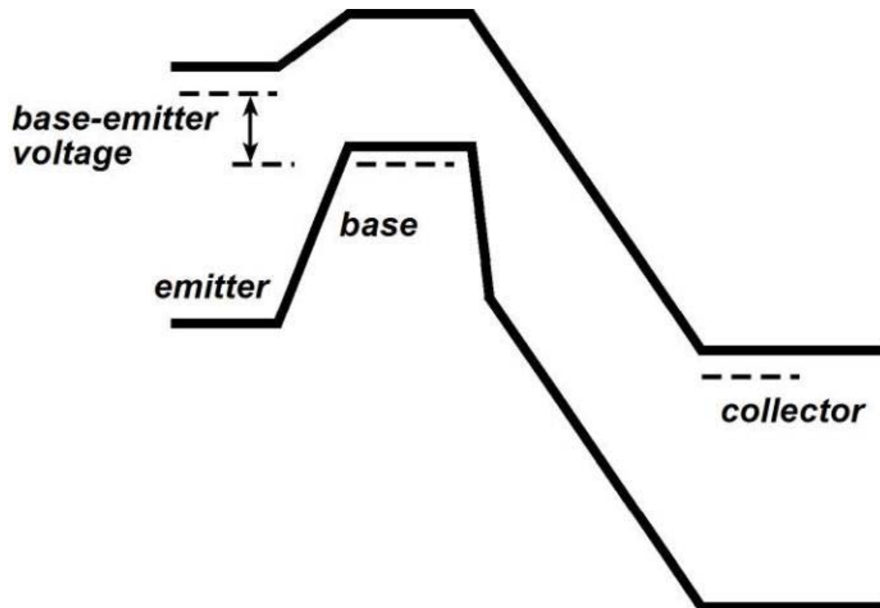
$$\delta V_{out} = -1 * \delta I_{plate} * R_L$$

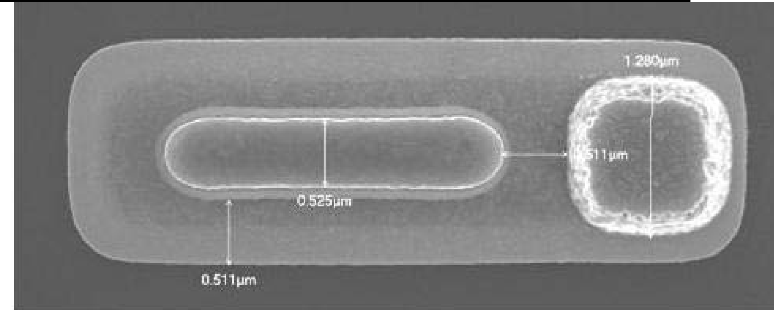
$$\text{Voltage Gain} = \frac{\partial V_{out}}{\partial V_{in}} = -\frac{\partial I_{plate}}{\partial V_{grid}} \times R_L = -g_m R_L$$

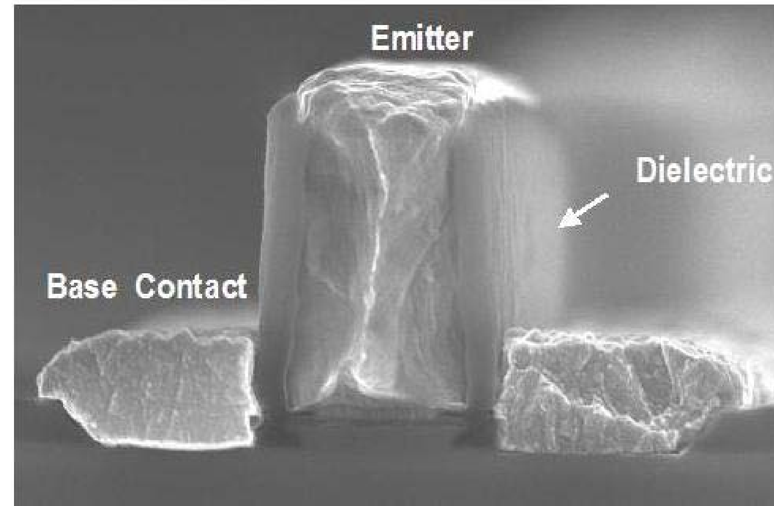
If we had time:
current gain, power gain
gain as a function of signal frequency

What Are Bipolar Transistors ?

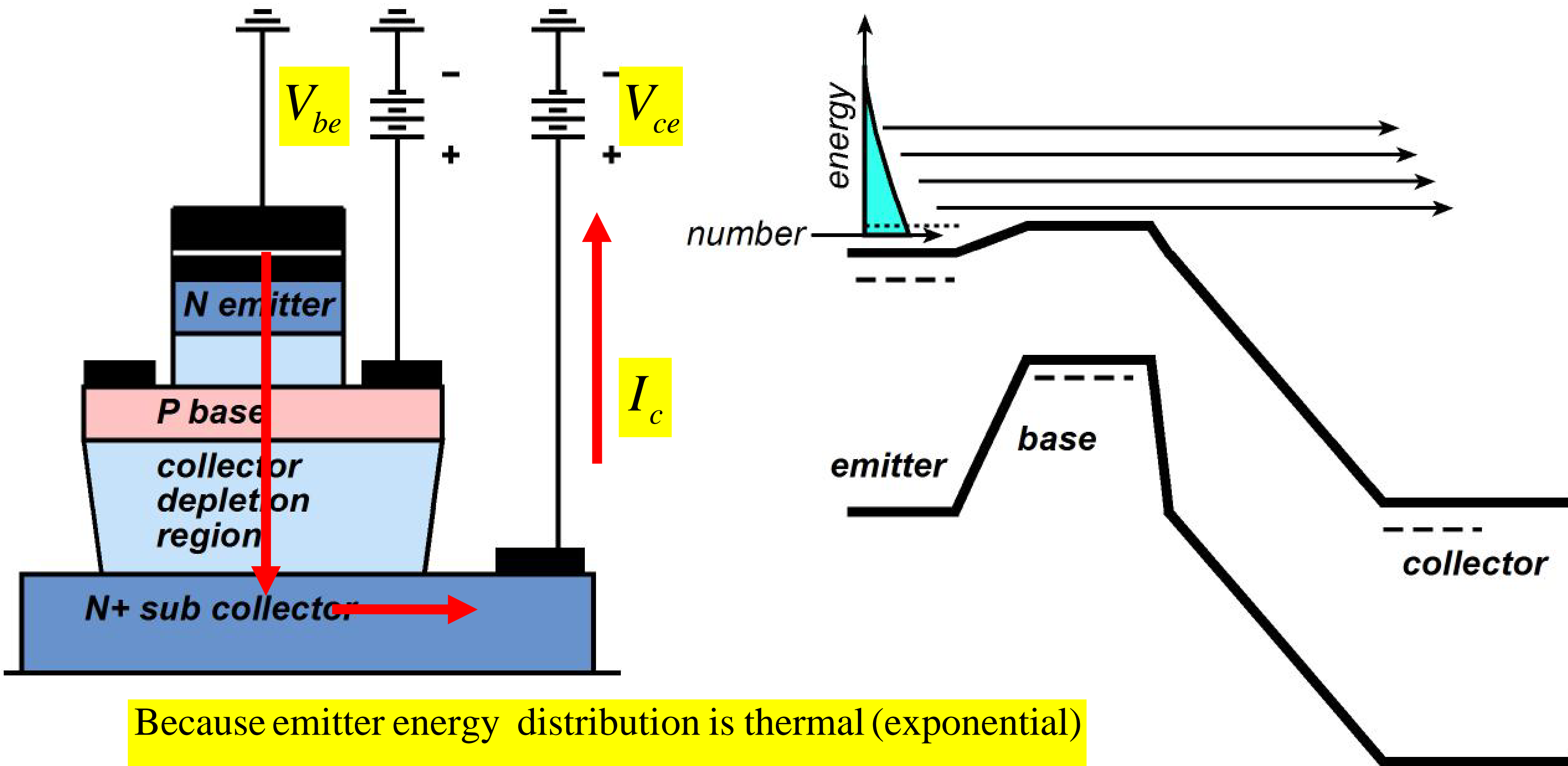








How Do Bipolar Transistors Work ?



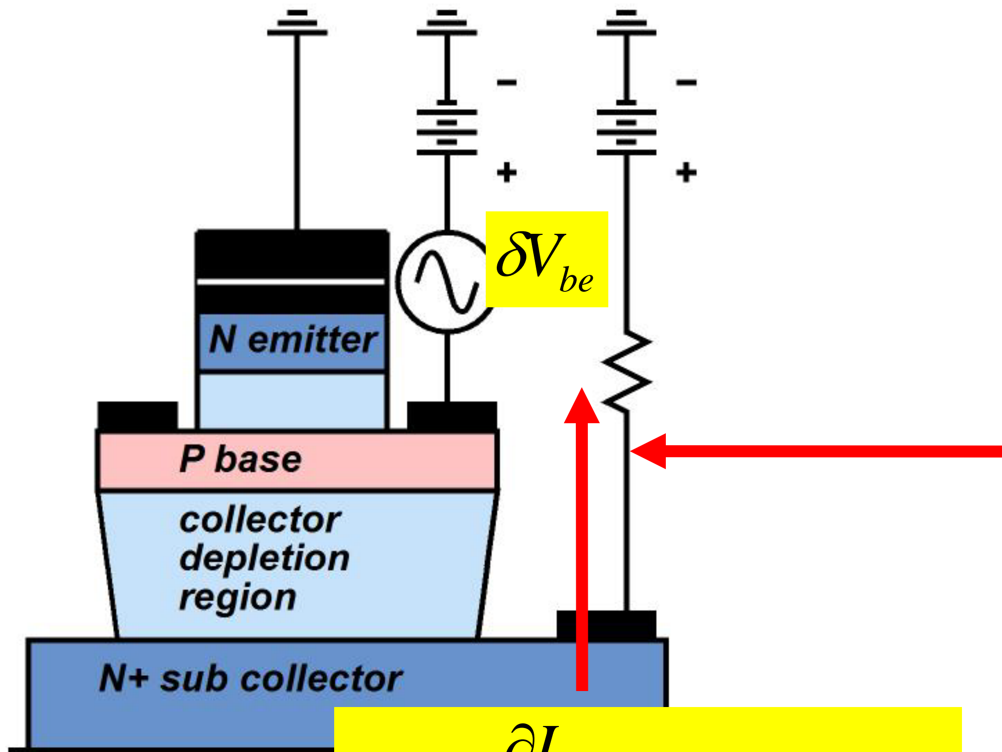
Because emitter energy distribution is thermal (exponential)

$I_c \propto \exp(qV_{be} / kT)$

Almost all electrons reaching base pass through it

$\rightarrow I_c$ varies little with collector voltage

How Do Bipolar Transistors Amplify Signals ?



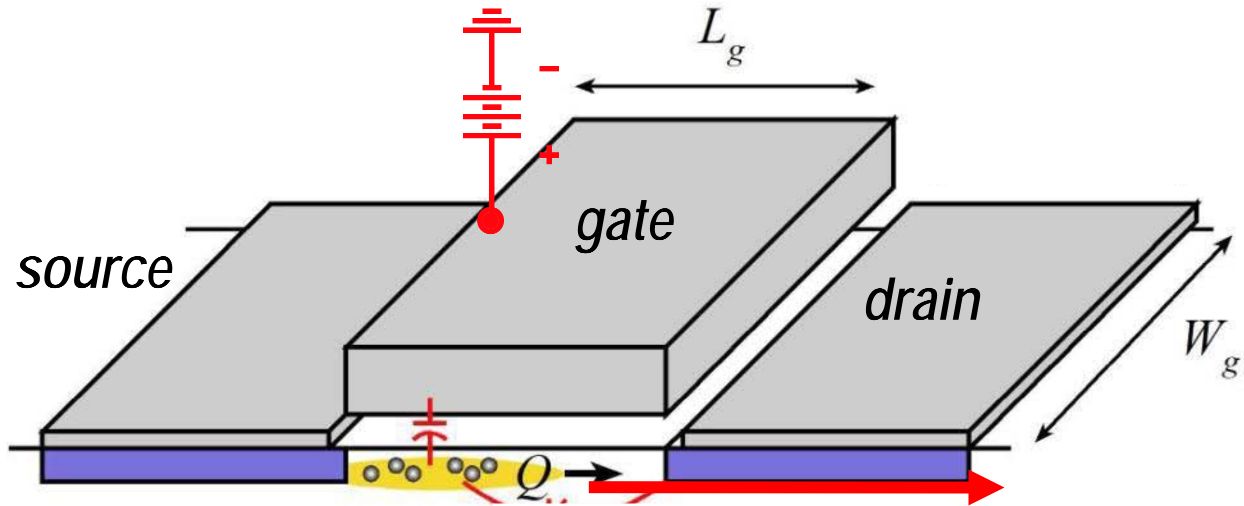
δV_{be}

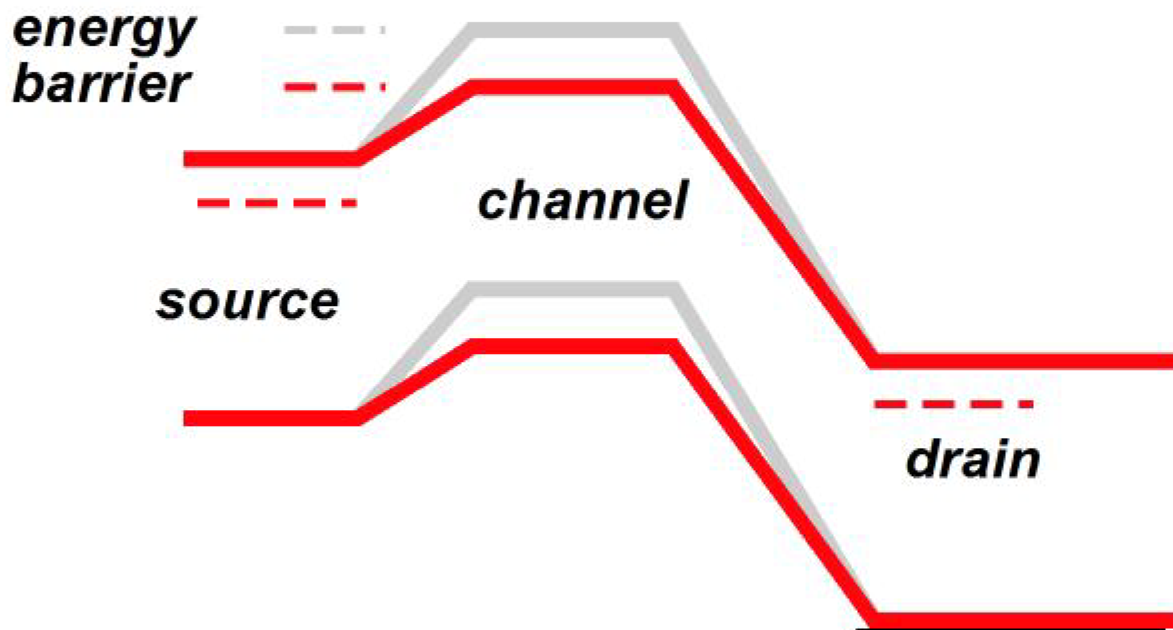
$$\delta V_{out} = \delta I_c \times R_L$$

$$\delta I_c = \frac{\partial I_c}{\partial V_{be}} \delta V_{be} = g_m \delta V_{be}$$

$$\text{Voltage gain} = \frac{\partial V_{out}}{\partial V_{in}} = -g_m R_L$$

How Do Field-Effect Transistors Work ?



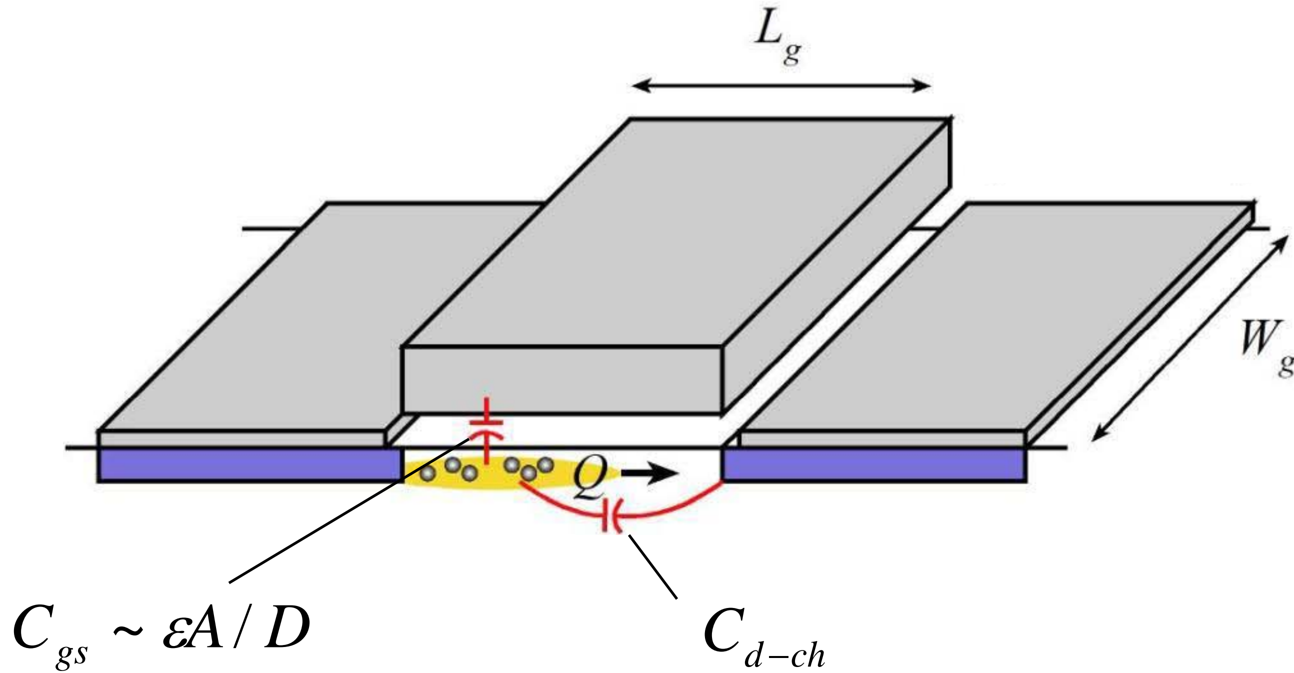


Positive Gate Voltage

→ **reduced energy barrier**

→ **increased drain current**

FETs: Computing Their Characteristics



$$I_d = Q / \tau \quad \text{where} \quad \tau = L_g / v_{electron}$$

$$\delta Q = C_{gs} \delta V_{gs} + C_{d-ch} \delta V_{ds}$$

$$\delta I_d = g_m \cdot \delta V_{gs} + G_{ds} \cdot \delta V_{ds} \quad \text{where} \quad g_m = C_{gs} / \tau \text{ and } G_{gd} = C_{d-ch} / \tau$$

FET Characteristics

L_g

W_g

Q

$C_{gs} \sim \varepsilon A / D$

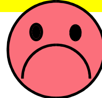
C_{d-ch}

I_D

increasing V_{GS}

V_{DS}

$$\delta I_d = g_m \cdot \delta V_{gs} + G_{ds} \cdot \delta V_{ds}$$

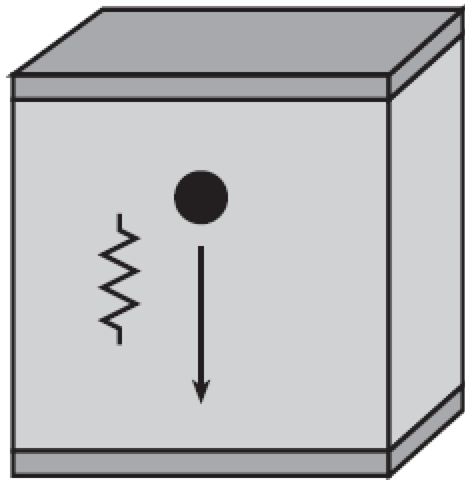


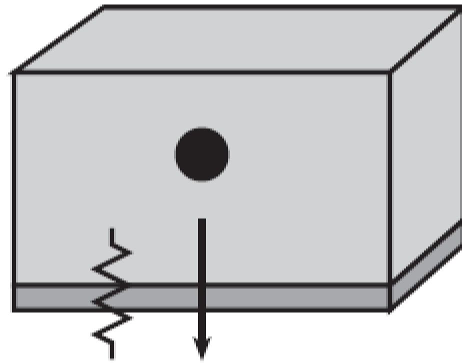
$g_m = C_{gs} / \tau$ $\quad G_{gd} = C_{d-ch} / \tau$ $\quad \tau = L_g / v_{electron}$

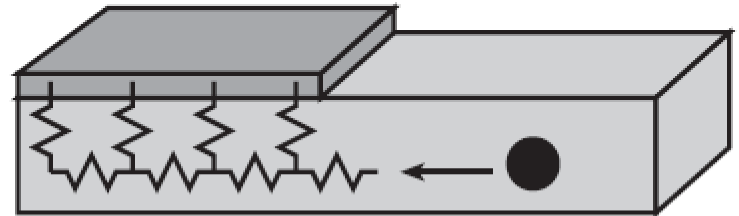
Tubes & Transistors

...what limits their frequency range ?

What Limits Semiconductor Device Bandwidth?



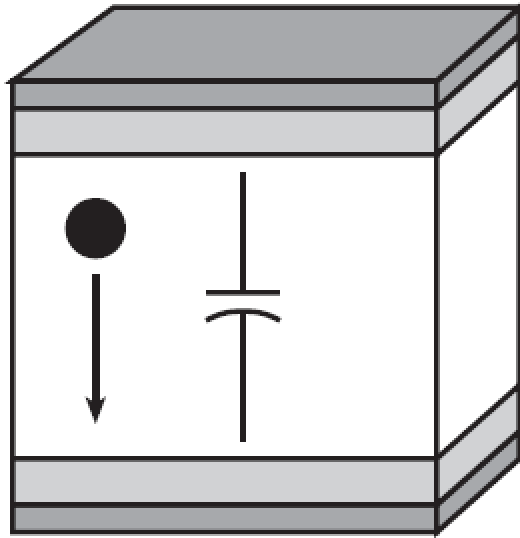




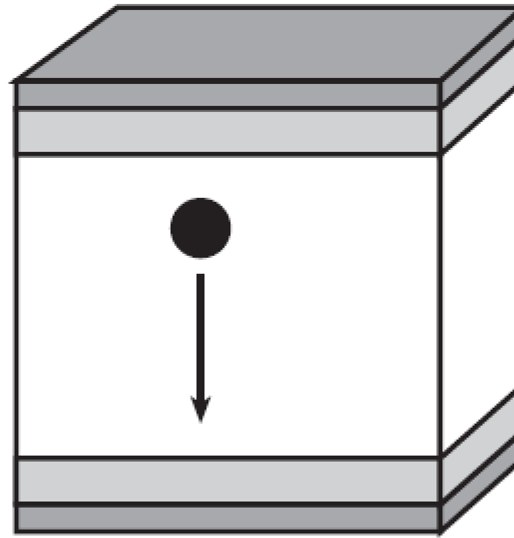
$R_{bulk} \approx \rho L / A$ $\quad$ $R_{contact} \approx \rho_{contact} / A$ $\quad$ $R_{cont,horiz} \propto 1/L$

- Bulk resistances
- Ohmic contact resistances
- Lateral contact access resistances
- These are for <u>undepleted</u> semiconductor layers

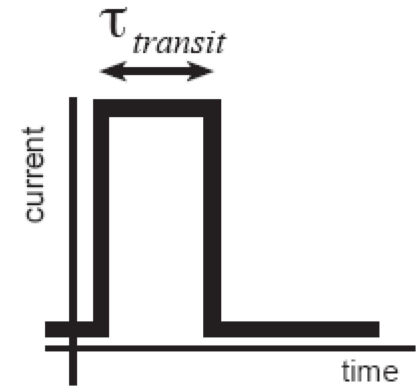
What Limits Semiconductor Device Bandwidth?



$C_{depl} = \varepsilon A / D$

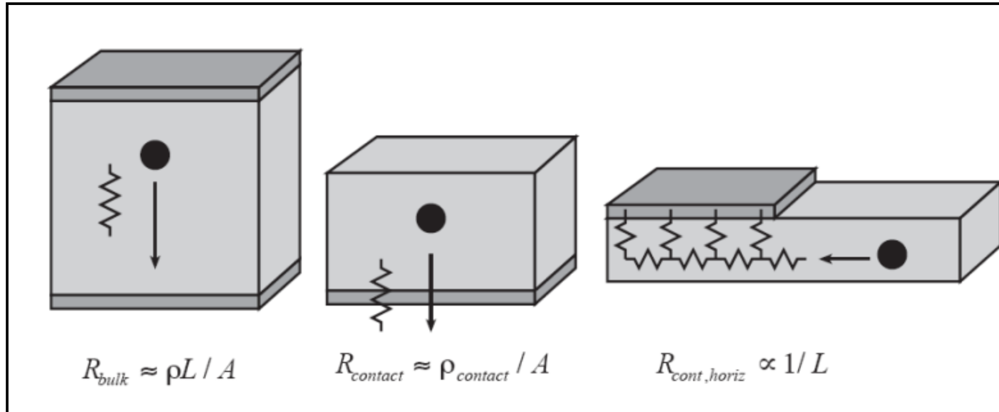


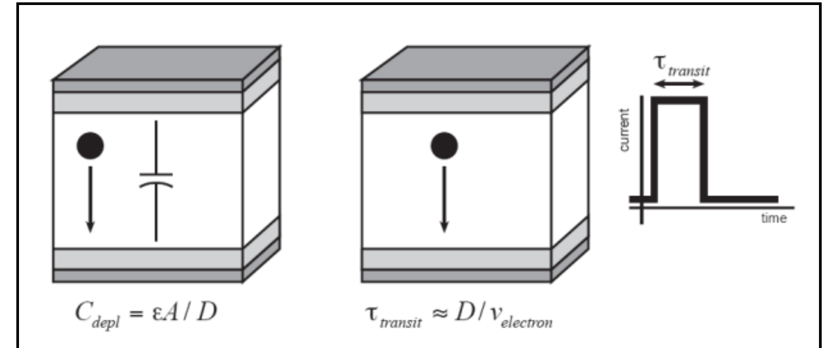
$\tau_{transit} \approx D / v_{electron}$



- Depletion layer capacitances
- Depletion layer transit times

Bandwidth Limits



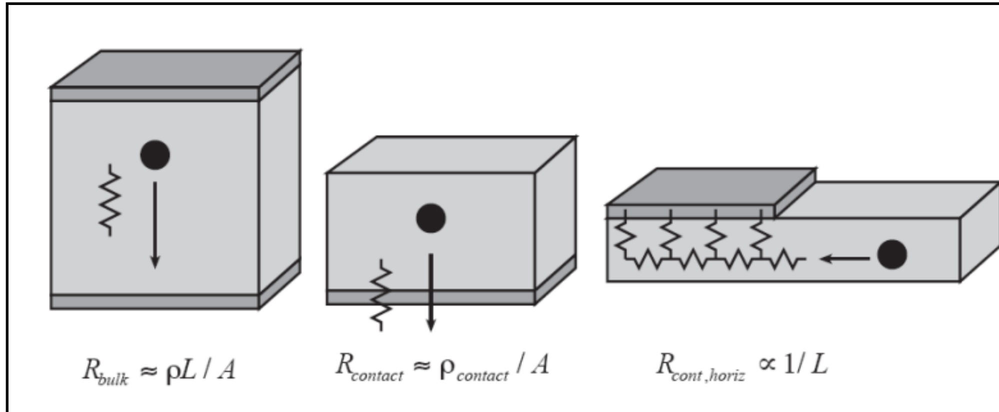


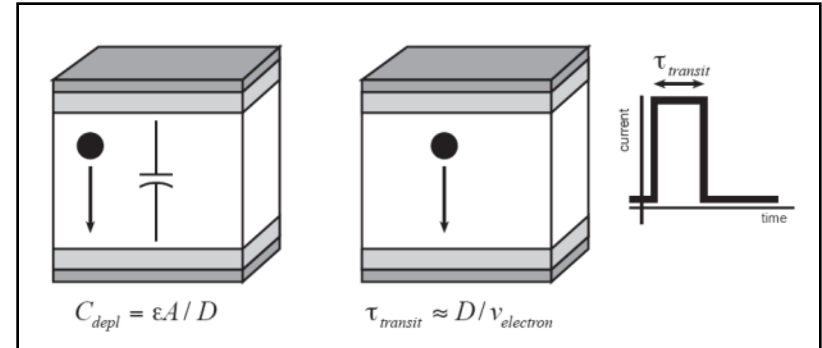
Frequency limits :

transit time : $\tau_{transit} = D / v_{electron}$

RC charging time : $\tau_{RC} = R_{access} C_{depletion}$

Bandwidth Limits



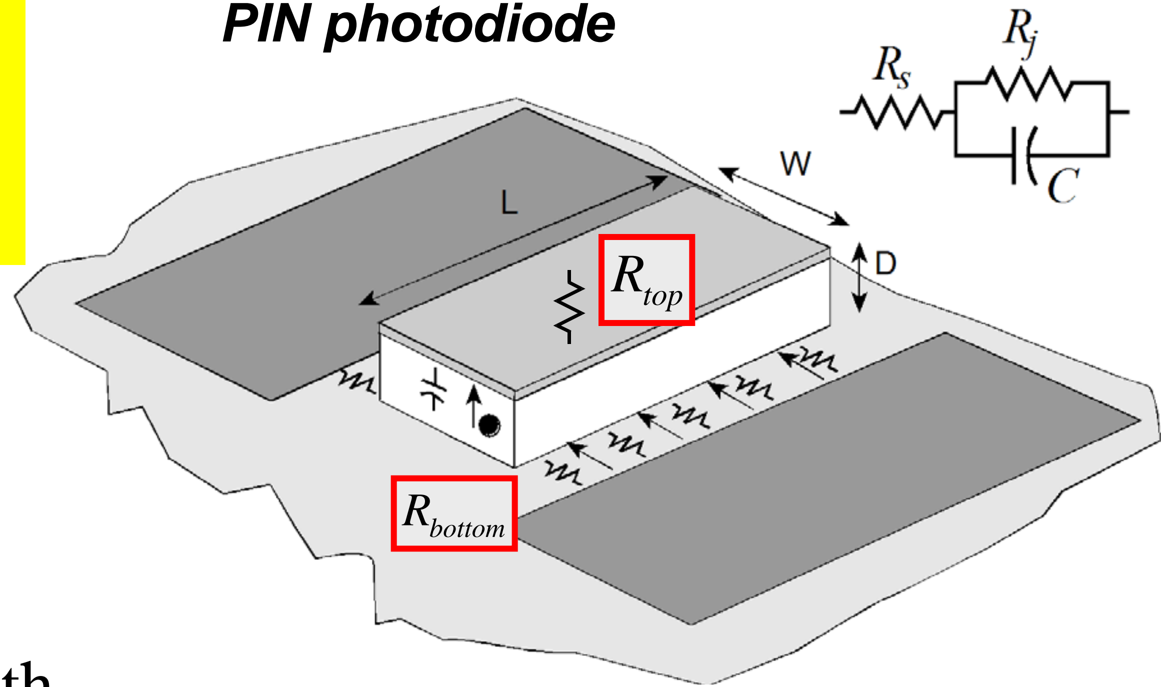


Frequency limits:

transit time: $\tau_{transit} = D / v_{electron}$

RC charging time: $\tau_{RC} = R_{access} C_{depletion}$

Frequency Limits and Scaling Laws of (most) Electron Devices



$$\tau \propto \text{thickness}$$

$$C \propto \text{area} / \text{thickness}$$

$$R_{top} \propto \rho_{contact} / \text{area}$$

$$R_{bottom} \propto 1 / \text{stripe length}$$

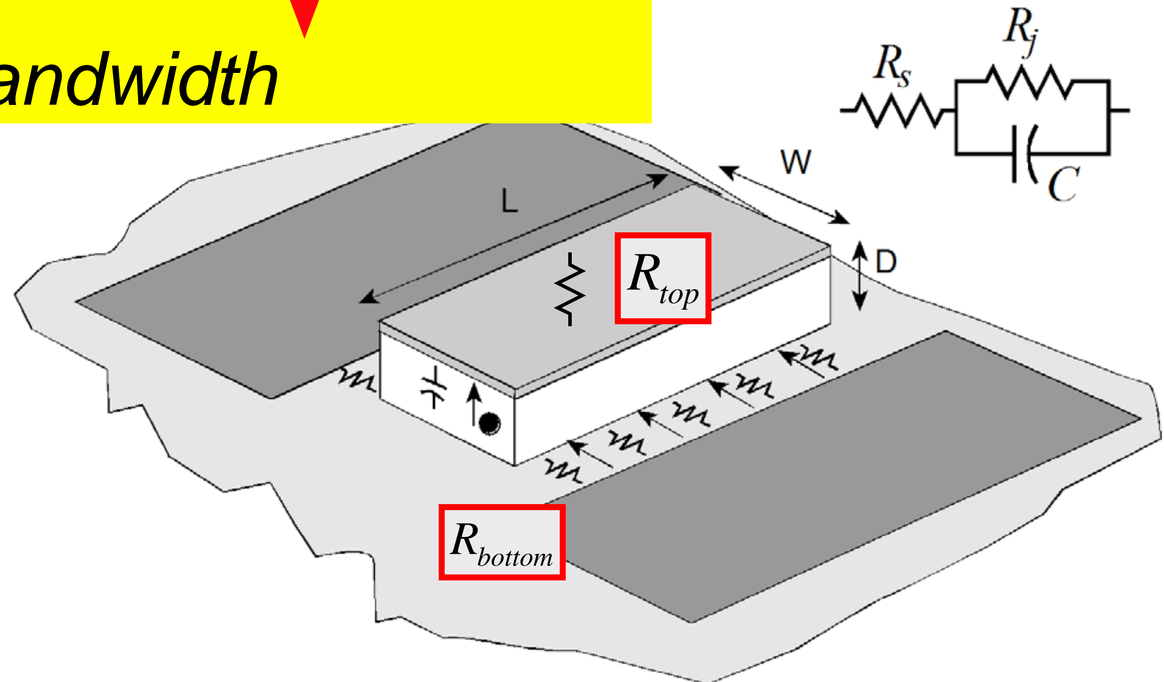
$$I_{\text{max, space-charge-limit}} \propto \text{area} / (\text{thickness})^2$$

$$\Delta T \propto \frac{\text{power}}{\text{length}} \times \log\left(\frac{\text{length}}{\text{width}}\right)$$

To double bandwidth,
reduce thicknesses 2:1
reduce width 4:1, keep constant length
current density has increased 4:1

resistance *capacitance* *transit time*

device bandwidth



applies to almost all semiconductor devices:

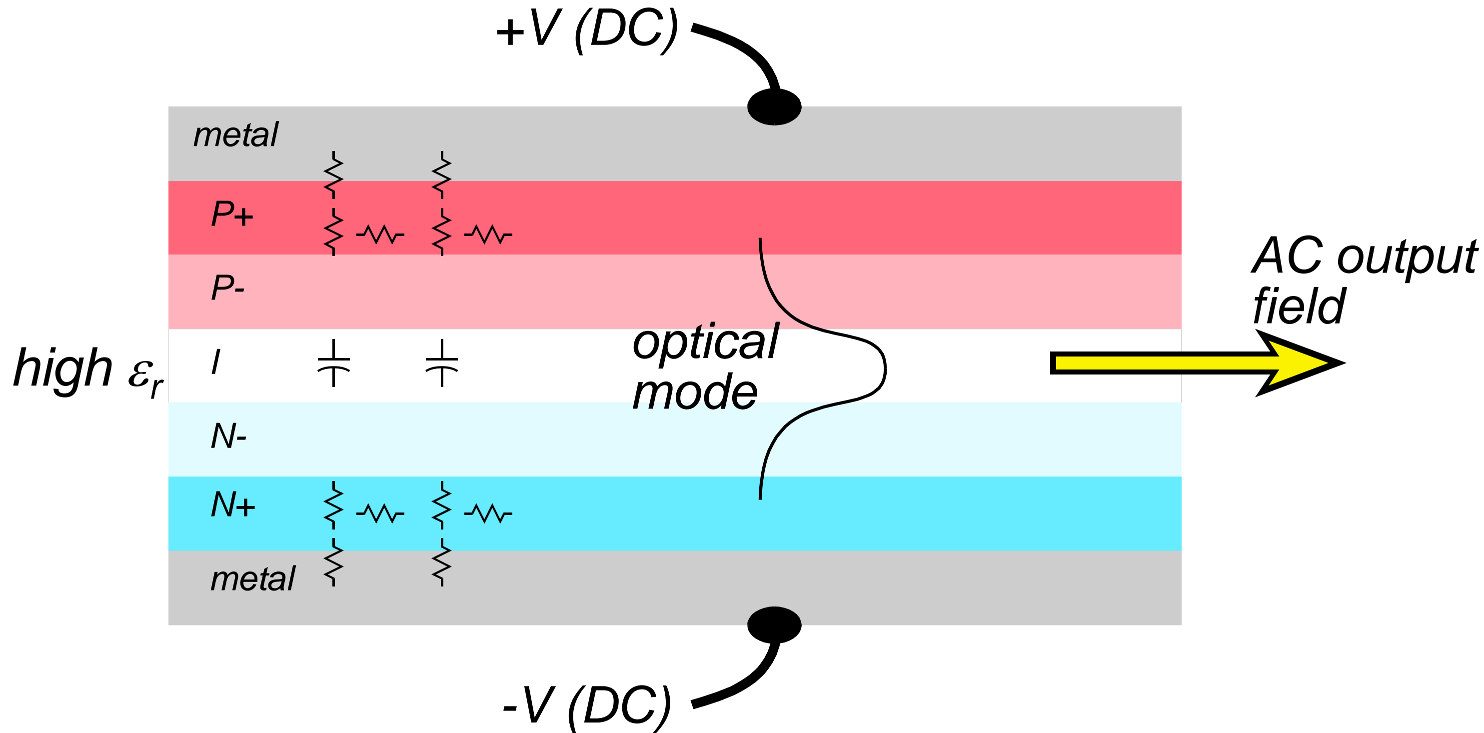
transistors: BJTs & HBTs, MOSFETS & HEMTs,
Schottky diodes, photodiodes, photo mixers, RTDs, ...

high current density,
low resistivity contacts,
epitaxial & lithographic scaling

THz semiconductor devices

FETs only: high $\varepsilon_r \varepsilon_o / D$ dielectrics

Why aren't semiconductor lasers R/C/τ limited ?



dielectric waveguide mode confines AC field away from resistive bulk and contact regions.

AC signal is not coupled through electrical contacts

dielectric mode confinement is harder at lower frequencies

Tubes & Transistors

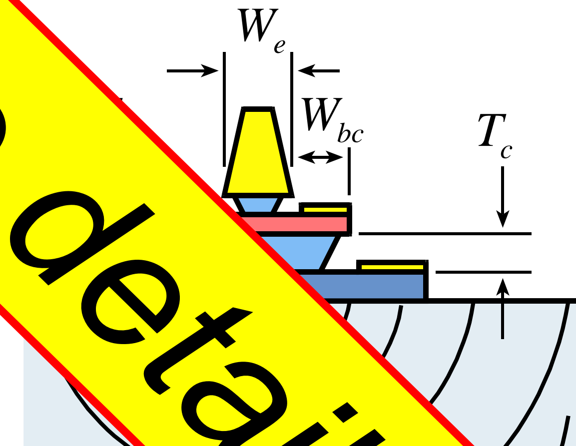
...increasing bandwidth
by scaling.

Bipolar Transistor scaling laws

Goal: double transistor bandwidth when used in ***an***

→ keep constant all resistances, voltages, currents

→ reduce 2:1 all capacitances and all transport delays



$$\tau_b = T_b^2 / 2D_n + T_b / v$$ → *thin base ~1.414:1*

$$\tau_c = T_c / 2v$$ → *thin collector 2:1*

$$C_{cb} \propto A_c / T_c$$ → *reduce junction areas 4:1*

$$R_{ex} = \rho_c / A_e$$ → *reduce emitter contact resistivity 4:1*

$$I_{c,Kirk} \propto A_e / T_c^2$$ *(current remains constant, as desired)*

$$\Delta T \cong \frac{P}{\pi K_{InP} L_E} \ln\left(\frac{L_e}{W_e}\right) + \frac{P}{\pi K_{InP} L_E}$$

need to reduce junction areas 4:1

reduce widths 2:1 & reduce length 2:1 → doubles ΔT

reducing widths 4:1, keep constant length → small ΔT increase ✓

$$R_{bb} \cong \frac{\rho_s W_e}{12 L_e} + \frac{\rho_s W_{bc}}{6 L_e} + \frac{\rho_c}{A_{contacts}}$$ → *reduce base contact resistivity 4:1*

reduce widths 2:1 & reduce length 2:1 → constant R_{bb} ✓

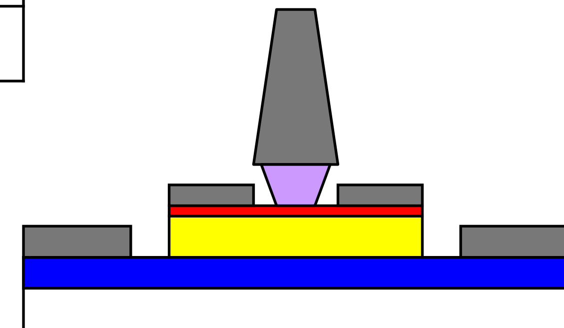
reducing widths 4:1, keep constant length → reduced R_{bb} ✓✓

Linewidths scale as the inverse square of bandwidth because thermal constraints dominate.

Bipolar Transistor Scaling Laws

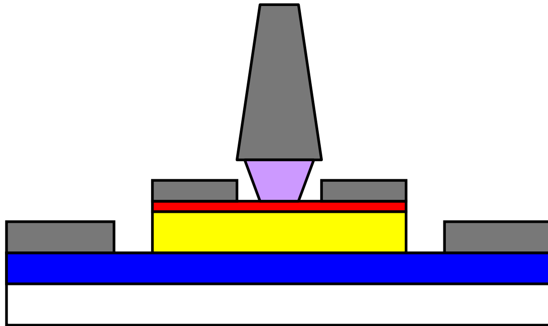
Changes required to double transistor bandwidth:

parameter	change
collector depletion layer thickness	decrease 2:1
base thickness	decrease 1.414:1
emitter junction width	decrease 4:1
collector junction width	decrease 4:1
emitter contact resistance	decrease 4:1
current density	increase 4:1
base contact resistivity	decrease 4:1



Linewidths scale as the inverse square of bandwidth because thermal constraints dominate.

Scaling challenges: What's hard ?

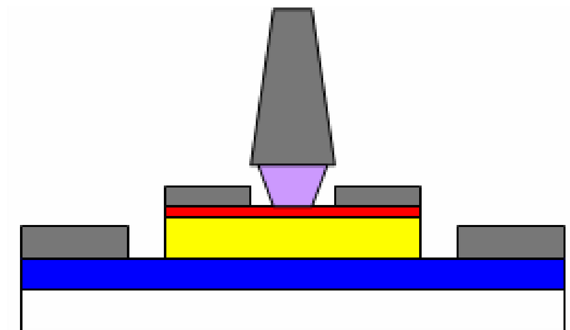


key device parameter	required change
collector depletion layer thickness	decrease 2:1
base thickness	decrease 1.414:1
emitter junction width	decrease 4:1
collector junction width	decrease 4:1
emitter resistance per unit emitter area	decrease 4:1
current density	increase 4:1
base contact resistivity (if contacts lie above collector junction)	decrease 4:1
base contact resistivity (if contacts do not lie above collector junction)	unchanged

Hard:
Thermal resistance (ICs)
Contact resistances
Yield in deep submicron processes
Reliability at very high current density

InP Bipolar Transistor Scaling Roadmap

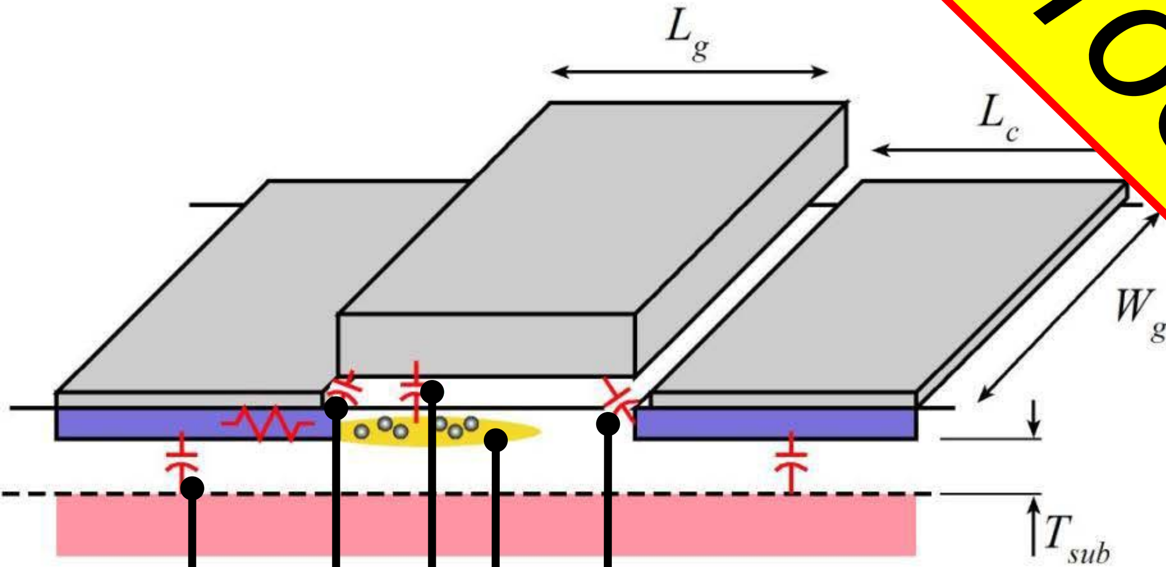
	industry	university →**industry**	university 2007-8	appears feasible	maybe
emitter	512	256	128	64	**32** nm width
	16	8	4	2	**1** $\Omega\cdot\mu m^2$ access ρ
base	300	175	120	60	**30** nm contact width,
	20	10	5	2.5	**1.25** $\Omega\cdot\mu m^2$ contact ρ
collector	150	106	75	53	**37.5** nm thick,
	4.5	9	18	36	**72** $mA/\mu m^2$ current density
	4.9	4	3.3	2.75	**2-2.5** V, breakdown
f_τ	370	520	730	1000	**1400** GHz
f_{max}	490	850	1300	2000	**2800** GHz
power amplifiers	245	430	660	1000	**1400** GHz
digital 2:1 divider	150	240	330	480	**660** GHz



Simple FET Scaling

Goal: double ... nsistor bandwidth when used in any circuit
→ red... all capacitances and all transport delays
... ant all resistances, voltages, currents

Too detailed



$$C_{gd}/W_g \sim \varepsilon$$

$$g_m/W_g \sim v\varepsilon/T_{ox}$$

$$C_{gs}/W_g \sim \varepsilon \cdot L_g/T_{ox}$$

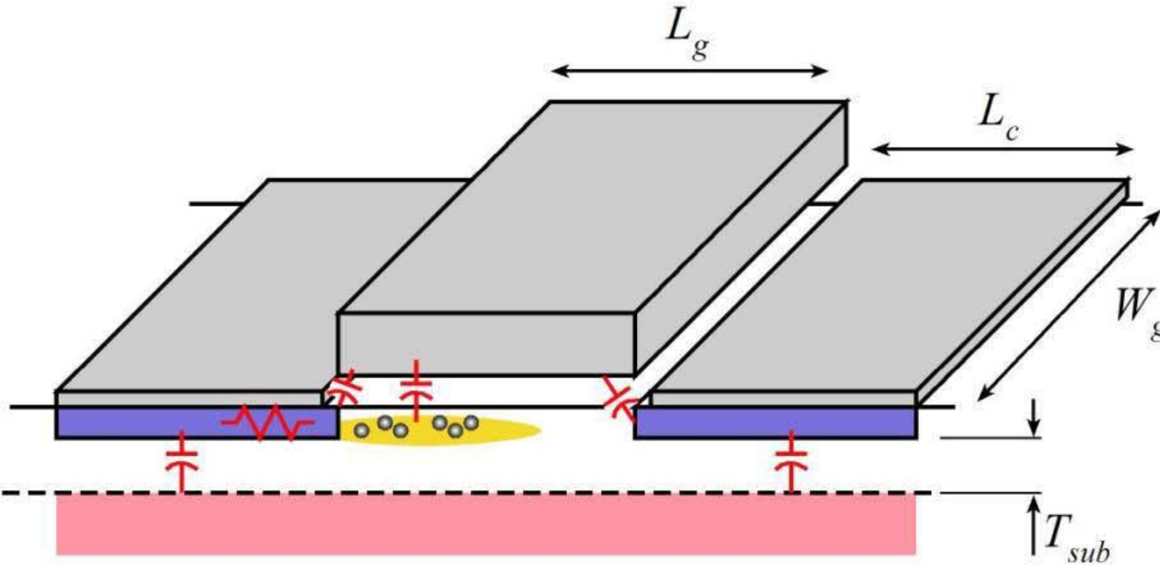
$$C_{gs,f}/W_g \sim \varepsilon$$

$$C_{sb}/W_g \sim \varepsilon \cdot L_c/T_{sub}$$

If T_{ox} cannot scale with gate length,
$C_{parasitic}/C_{gs}$ increases, ✕
g_m/W_g does not increase
hence $C_{parasitic}/g_m$ does not scale ✕

Simple FET Scaling

Goal: *double transistor bandwidth when used in any circuit*
→ reduce 2:1 all capacitances and all transport delays
→ keep constant all resistances, voltages, currents

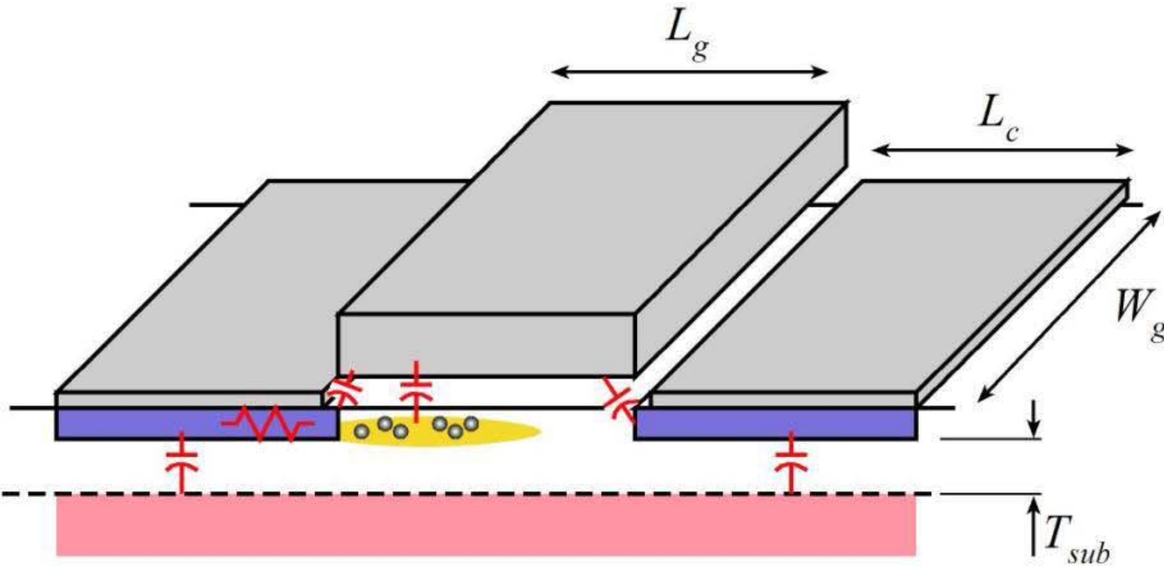


decrease gate length 2:1 (easy?)

decrease contact resistivities 4:1 (hard)

Increase gate capacitance/area 2:1 (very hard)
  tunneling limits in thin insulators
  upper limit on C/A from δQ/δV of semiconductor itself

Scaling challenges: What's hard ?



Hard:
Contact resistances
Gate capacitance density ($\varepsilon_r\varepsilon_o/D$)

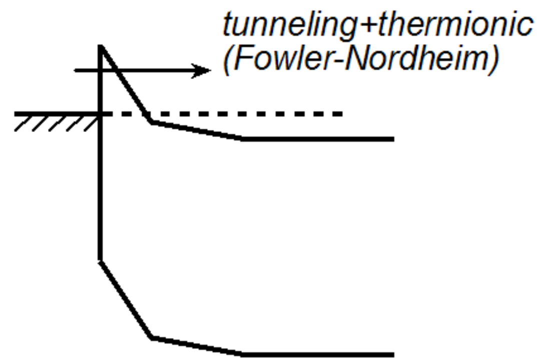
nm / THz Transistors

So...what are we working on ?

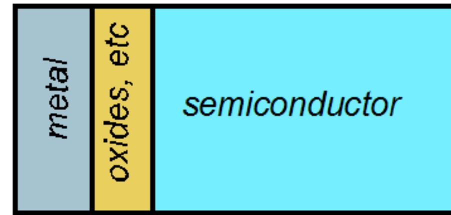
Bipolar Transistors → THz ICs

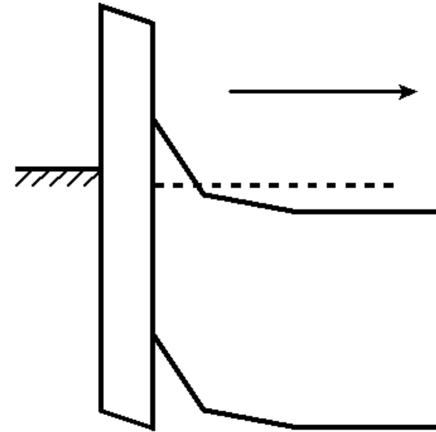
Conventional ex-situ contacts are a mess

textbook contact

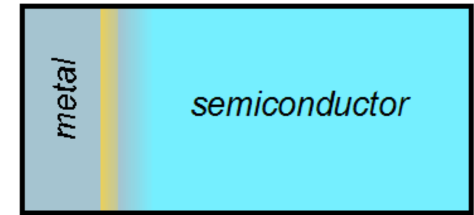


with surface oxide





with metal penetration



*Interface barrier → **resistance***

*Further intermixing during high-current operation → **poor reliability***

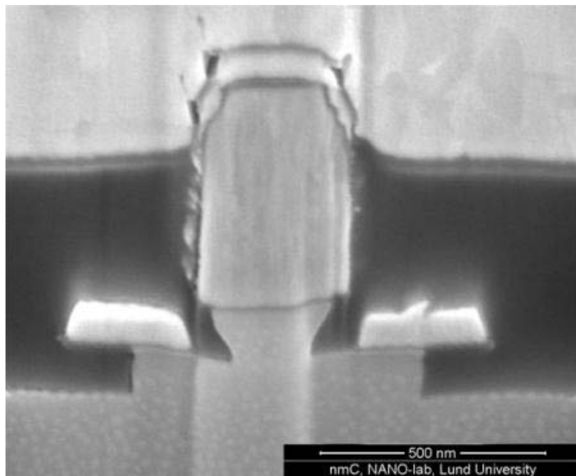
So, we are working on Forming contacts in ultra-high vacuum, perhaps even by MBE

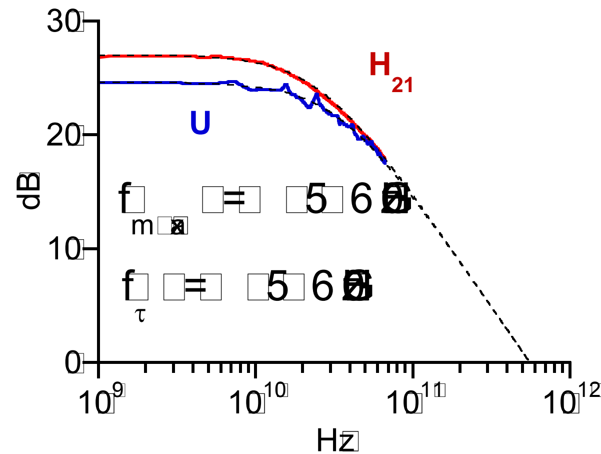
Current UCSB 250 /125 nm Mesa HBT process

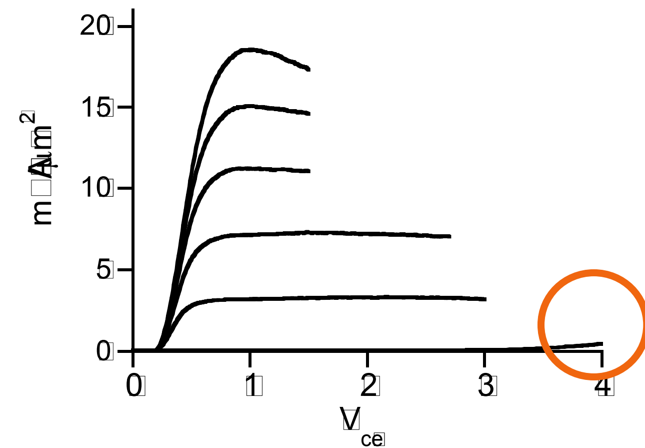
Litho	pattern metal	sidewall	dry etch	wet etch

SiO2
TiW
Ti
InGaAs n++
InP n
InGaAs p++ Base

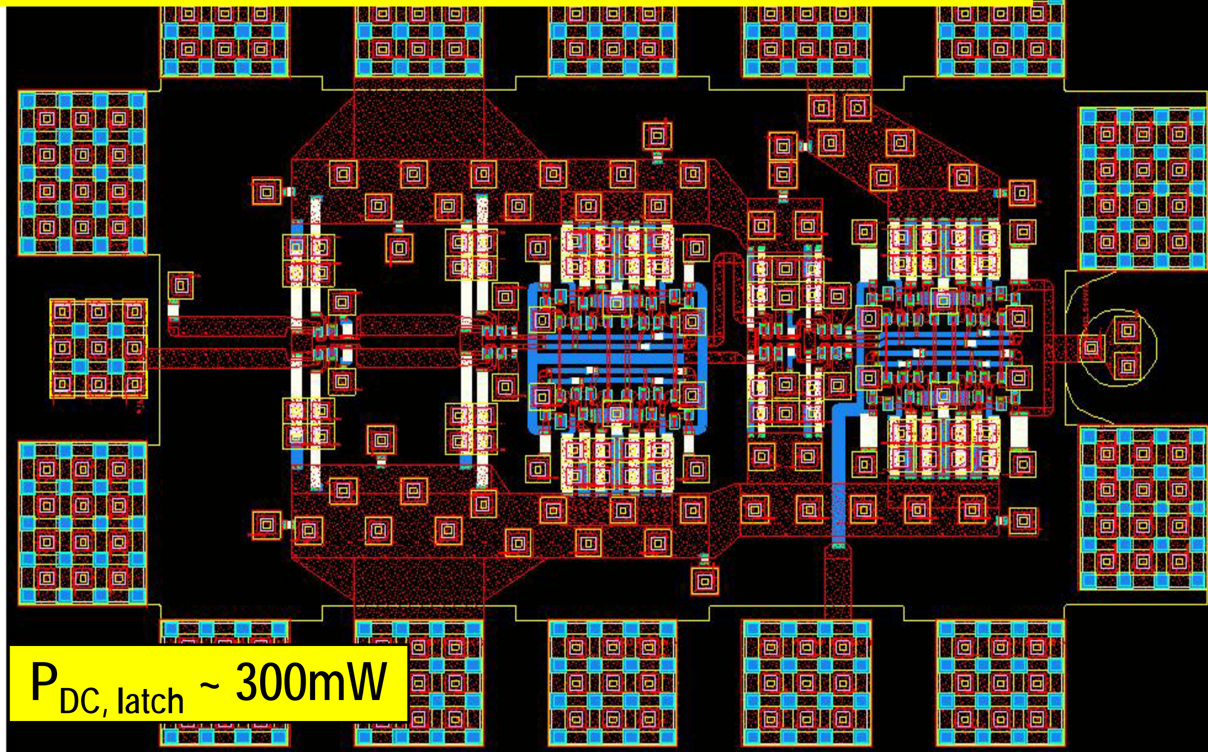
BHF

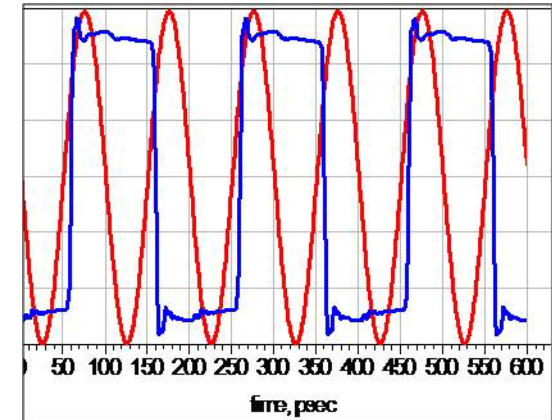




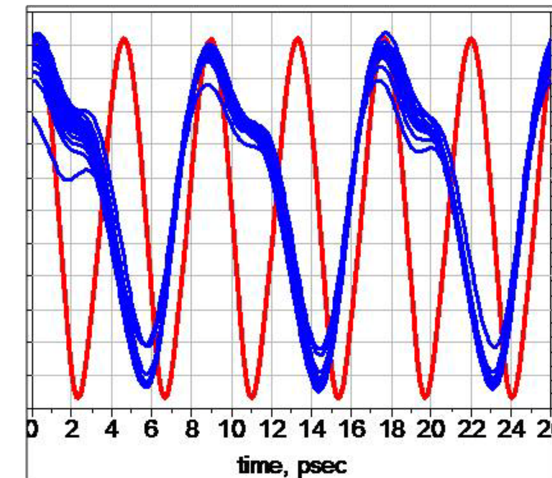


200 GHz Digital IC designs : 250 nm HBT



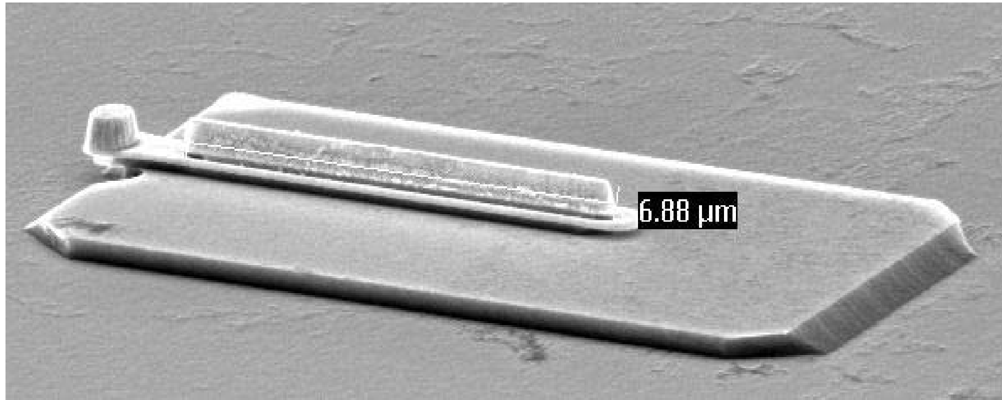


Simulation:
f_{clk} = 10GHz, f_{out} = 5GHz

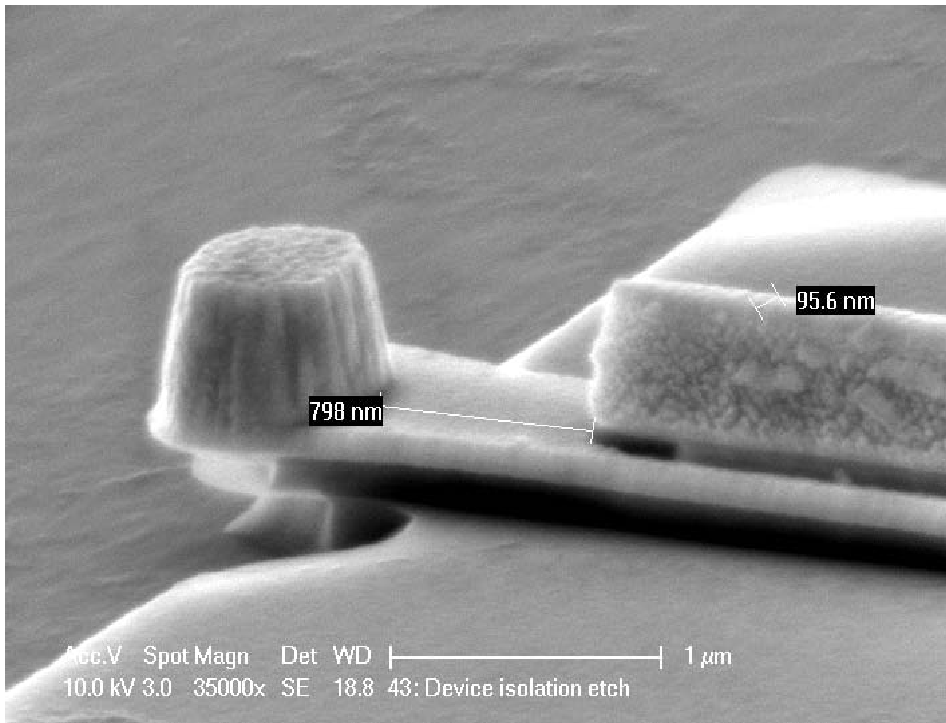


Simulation
f_{clk} = 230GHz, f_{out} = 115GHz

We Are Working on 128-nm HBTs

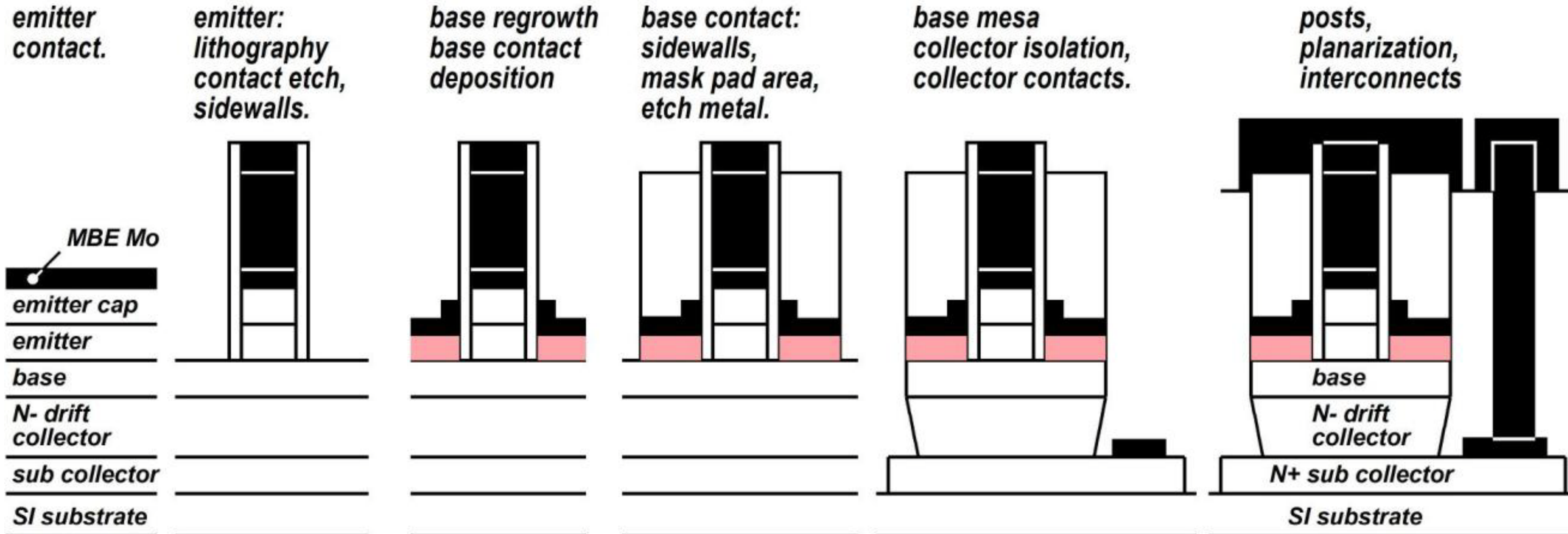


128 nm process runs seem to be getting close.



We hope to get 1.2 THz bandwidths from these

Next-Generation HBT Process Flow



Key Process steps (base & collector contacts)

*by MBE → **ultra low resistivity contacts ?***

*→ **2**-3 THz bandwidths ??*

nm / THz Transistors

So...what are we working on ?

III-V MOSFETs for VLSI

Why Develop III-V MOSFETs ?

Silicon MOSFETs continue to scale...

...22 nm is feasible in production (or so the Si industry tells us...)

...16 nm ? -- it is not yet clear

If we can't make MOSFETs yet smaller,

instead move the electrons faster:

$$I_d / W_g = q n_s v \qquad I_d / Q_{transit} = v / L_g$$

III-V materials→ **lower m***→ **higher velocities**

Serious challenges:

High-K dielectrics on InGaAs channels,

InGaAs growth on Si

True MOSFET fabrication processes

Designing small FETs which use big (low m*) electrons

Highly Scaled MOSFETs: What Are Our Goals ?

Low off-state current (10 nA/μm) for low static dissipation

→ minimum subthreshold slope→ minimum L_g / T_{ox}

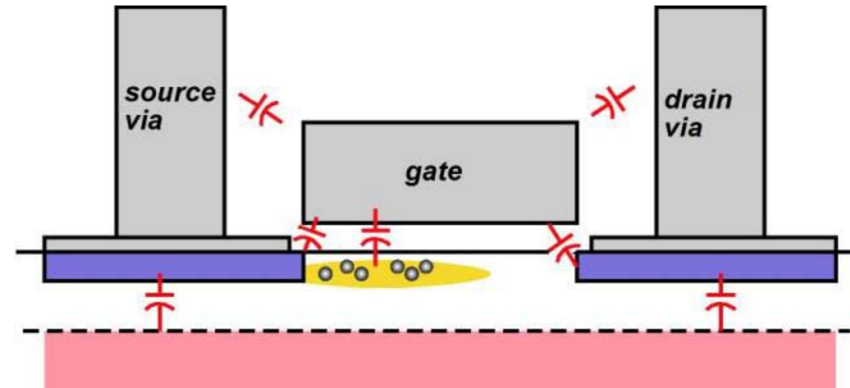
low gate tunneling, low band-band tunneling

Low delay $C_{FET}\,\Delta V/I_d$ in gates where transistor capacitances dominate.

Parasitic capacitances are 0.5-1.0 fF/μm

→ while low C_{gs} is good,

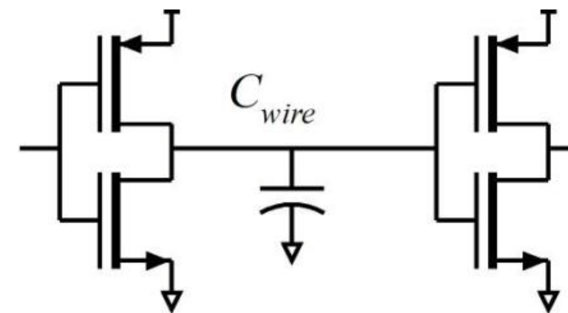
high I_d is much better



Low delay $C_{wire}\,\Delta V/I_d$ in gates where wiring capacitances dominate.

large FET footprint → long wires between gates

→ need high I_d / W_g ; target ~6 mA/μm



Very Rough Projections From Simple Ballistic Theory

22 nm gate length

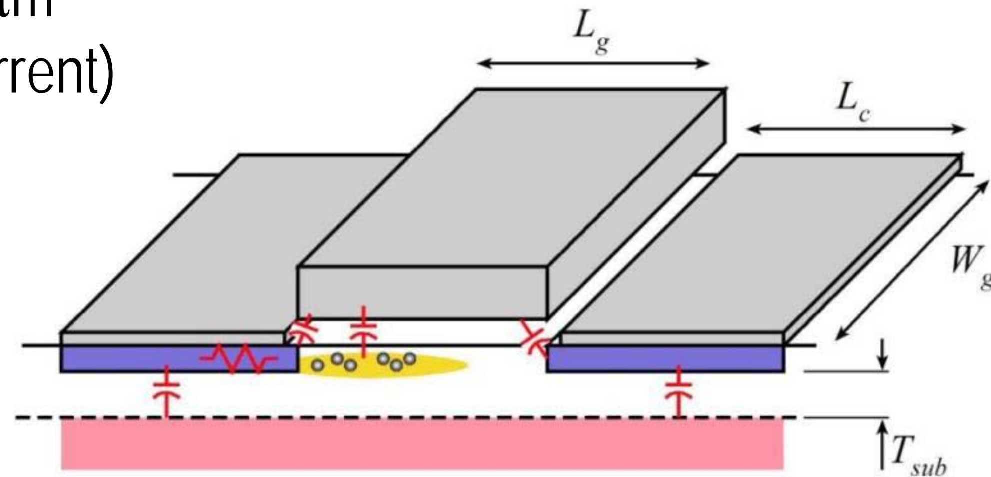
0.5-1.0 fF/μm parasitic capacitances

Channel	EOT	drive current (700 mV overdrive)	intrinsic (transport) gate capacitance
InGaAs	1 nm	6 mA/μm	0.2 fF/μm
InGaAs	1/2 nm	8 mA/μm	0.25 fF/μm
Si	1 nm	2-4 mA/μm	0.7 fF/μm
Si	1/2 nm	5-7 mA/μm	1.4 fF/μm

InGaAs has much less gate capacitance
1 nm EOT → InGaAs gives much more drive current
1/2 nm EOT → InGaAs & Si have similar drive current
InGaAs channel→ little benefit for sub-22-nm gate lengths

Implications for Our Device Designs

Device
- drive current > 5 mA/μm at ~700 mV overdrive
- inversion carrier concentration: 10^{13} /cm^2
- off-state current must be < 10 nA/μm
- Low CV/I delays (will get if high current)



Dielectric:
- EOT < 1 nm, 0.6 nm preferable
- interface D_{it} < about 5*10^{11} /cm^2

Channel :
- high-mobility InGaAs <5 nm thick
- mobility > 1000 cm^2/V-s at 5 nm thickness, 10^{13} /cm^2

S/D access resistance:
- <10 Ohm-μm resistivity, >2*10^{13} /cm^2 carrier density, < 5 nm thick

Galileo, Elephants, & Fast Nano-Devices

Semiconductor Device Scaling

Scaling is the key to success of
- CMOS VLSI,
- microwave/ mm-wave III-V electronics

Scaling will take III-V transistors well in to the THz

Scaling limits are at the surfaces
- contact resistivities
- dielectric capacitance densities

Scaling limits also come from heat
- current densities
- device thermal resistance
- IC thermal resistance

Scaling

Changing the scale changes:

Perimeter / area / volume ratios,
which changes characteristic times, strength / weight ratios...
electrons move in femtoseconds, Galaxies in aeons

The dominant physics changes with scale, too:
A human feels the Coulomb force (as mechanics),
Galaxies mostly driven by gravity