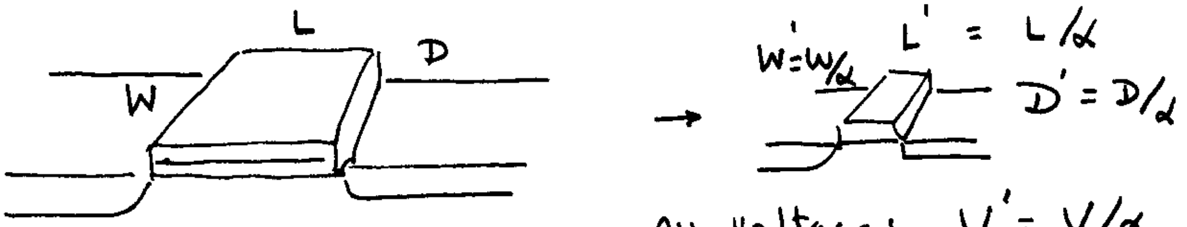
LECTURE #15 NOVEMBER 9

- Collect Proj. Assign. #2
- <u>LAB/PROJ</u>: Use only Boxes at right ∡'s. Later software won't support more.
  Also: We have priority in Lab > 3P.M.
  POINT OUT $ COSTS: ONCE DES RIGHT, STAMP OUT LIKE COOKIES AT ~ $1 OR SO APIECE.
- <u>TODAY</u>: CONTINUE SCALING.
  DISCUSS LIMITING FACTORS.

- <u>SUMMARIZE SCALING SO FAR</u>:



All Voltages $V' = V/\alpha$

Using this simple scaling by α of all dimensions including vertical, and all voltages (VDD, V_{TH}, V_{DEP}) we found:

$\boxed{\tau' = \tau/\alpha}$; $\boxed{I_{ds}' = I_{ds}/\alpha}$; <u>PROBLEM ENCOUNTERED</u>!

[ex.: EQ 1: $\tau = L^2/\mu V_{ds} \therefore \downarrow$]

But # of devices per unit area goes up by α^2
So mean current density: <u>CURRENT INTO AREA OF CHIP GOES UP BY α</u>. This is real problem, since in our scaling wires would get thinner. Current limited wires would either have to have aspect ratios increase by α^2 (which can't go on for long) <u>or</u> get proportionally wider.

$\boxed{C_g' = C_g/\alpha}$; But <u>capacitances</u> in general <u>scale up</u> by α if given in terms of $C/\mu m^2$ since vertical dimensions shrinking (oxide getting thinner).

Resistances/□ scale up by α since lines getting thinner.

(Except note that we can't really do that with curr. lim. metal.)

However, if calculate it out, find that R/□ of FETS will stay about the same.

So another problem: R/□ of poly, diff getting proportionally larger while R/□ FET staying same. This is aggravated by crystal clumping in POLY --- makes effect worse as scale down.

DC Power Dissipation: Per Device $P_{dc} = I \cdot V$

$$\boxed{P_{dc}' = \frac{P_{dc}}{\alpha^2}}$$

; # Dev. / Area goes up as α^2

So (whew!) Power dissipation/unit area stays approx constant.

We discussed power dissipation limits: Is much more diff. to pin down to single constraint as in current density in wires. Dependent on next level context

Tabulated:

< 1 W/cm^2	no prob.
2 W/cm^2	begin to need reasonable heat sinking
4 W/cm^2	
> 8 W/cm^2	need way to remove heat: forced cooling

We examined ≈ worst case in our methodology:

Large array of shift registers are in Fig 86 CH 4.

Found power/unit area ≈ 10 W/cm^2.

BUT NOTED THAT WE USE PASS-T LOGIC BETWEEN THESE AND USUALLY DON'T USE SUCH SHORT PULLUPS, etc.

SO NORMALLY WE DON'T NEED TO WORRY. BUT SHOULD CALCULATE IF IN DOUBT. TRY TO STAY < 2 W/cm^2 IF CAN.
(cover large areas of full chip)

Scaling of
SWITCHING POWER: The drivers which operate pass gates, charging & discharging capacitances dissipate switching power.

The power is dissipated at the drivers, but we calculate the amount based on the capacitances & voltages & clock period:

P_{sw} = energy stored on capacitance divided by the clock period or time between successive chargings/discharging.

But $T^c \propto \tau$. Thus, $P_{sw} \propto \dfrac{CV^2}{T}$; $\tau \propto \dfrac{L^2}{V}$

$$\therefore P'_{sw} \propto \frac{WL}{D} \cdot V^2 \cdot \frac{V}{L^2} = \frac{WV^3}{DL} \quad \therefore \boxed{P'_{sw} = \frac{P_{sw}}{\alpha^2}}$$

So: Since P_{sw} per device goes down by α^2, and # dev./area goes up by α^2, the switching power also stays constant per unit area as we scale things down.

NOTE: AVERAGE DC POWER DISS. IN MOST SYSTEMS CAN BE APPROXIMATED BY ADDING TOTAL P_{sw} TO 1/2 DC POWER RESULTING IF ALL LEVEL RESTORING LOGIC WERE TURNED ON.

SWITCHING ENERGY

We've noted before that there are various ways to trade off power vs delays. We can often use less power if we can tolerate longer delays, and vice-versa. This can be done by building it into the design, or sometimes can be controlled dynamically. This reflects an important metric of device performance: <u>SWITCHING ENERGY per DEVICE</u>

E_{sw} = power consumed by device at max clock freq, multiplied by the delay: i.e., it is a "power × delay" product.

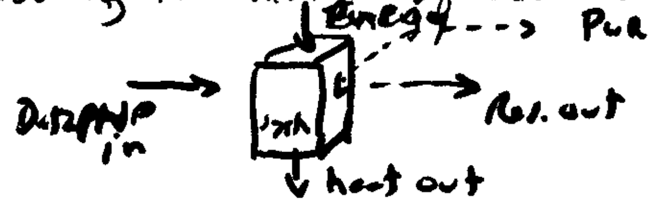
<u>Rationale</u>: In a sense, to do any computation, we must, <u>at minimum</u>, switch a large collection of switches, (Think of toggle switches) switching them in a particular order, and some number of times.

Switches have some switching energy which measures the work done to throw the switch. We often have the option (by design or control) to choose to put the energy into a system slowly, (and thus less power) taking longer for a calculation. Or - put it in faster, flipping the switches faster. (<u>SEE CHAP</u> 9)

However, there are usually constraints on both speed and power, and these limit ultimate performance:

No matter how much power we put in, we can't reduce delays below the minimum value of τ.

Also, if we put in too much power, we may reach a power density limit in a particular design even before reaching the fastest speeds.



$$\text{Min Pwr} \times \text{Time} = \underset{\text{of comp.}}{\text{Min Energy}} = \#\text{switchings} \times \underset{E_{sw}/\text{dev}}{\text{device}}$$

HOW DOES SWITCHING ENERGY SCALE?

Our basic FET switches have $E_{sw} \propto CV^2$

and ∴ $\boxed{E'_{sw} = \frac{E_{sw}}{\alpha^3}}$

so this crucial metric of device performance scales incredibly favorably. This is why scaling down sizes is so important.

Summary So Far: Suppose we

Scale down an entire system by $\alpha = 10$.

> Resulting system will have 100X as many devices/unit area.
(Or take only 1/100 th as many chips)
> Power Density remains constant.

> All voltages reduced by factor of 10
> Current/Area (circuit) increased by factor of 10

> Time delay/stage decreased by factor of 10
> Power-Delay Product of Devices decreased by " " 1000

This is very attractive scaling except for the current density problem. The delivery of the required average dc current presents an important obstacle to scaling. Even in today's systems, many wires are operated at near there current limit. So, wires must become relatively wider, or have much higher aspect ratios, or both.

(& Don't Forget the problems of Poly, Diff res/□ rel to FET)

Forgetting possible design rule changes, wire problems:
[CAN WE SIMPLY SCALE DOWN WHOLE DESIGN? Yes but:]
CONSIDER

- Delays to Outside World: (Read SPACE vs TIME in CH 1)

What is effect of scaling on output driver design & delays? We can't just scale them down: the outside world stays big. Remember the result in chap 1:

Min Delay when use factor of e, and $N = \ln Y = \ln \frac{C_L}{C_{g_{min}}}$

In this case: Min Tot Delay $\approx \tau e \ln\left[\frac{C_L}{C_g}\right]$

Now, scale everything down by α, including Voltages. (This we do scale even in the external world)

$\tau' = \tau/\alpha$; $C_g' = C_g/\alpha$; $\therefore$ $Y' = \alpha Y$

DERIVE ↓ ↱ *

$$\therefore \boxed{t'_{min} = t_{min} \cdot \frac{1}{\alpha}\left[1 + \frac{\ln \alpha}{\ln Y}\right]}$$

below for deriv. if necessary

So, as inverters get smaller, more stages are required to obtain minimum offchip delay.

The relative delay to outside world increases, But the absolute delay decreases!

ALSO: At Least Driver Designs must change & can't be just scaled down like rest of system (cont.)

Derive *:

$$t_{min} = \tau e \ln Y$$

$$t'_{min} = \tau' e \ln Y' = \frac{\tau}{\alpha} e \ln(\alpha \cdot Y) = \frac{\tau}{\alpha} e\left[\ln \alpha + \ln Y\right]$$

$$t'_{min} = \frac{\tau e \ln Y}{\alpha}\left[1 + \frac{\ln \alpha}{\ln Y}\right] = \frac{t_{min}}{\alpha}\left[1 + \frac{\ln \alpha}{\ln Y}\right]$$

ALSO, **BRIEFLY** MENTION BIPOLAR TTL, I^2L, OTHER MOS: CMOS

SO, SCALING PRODUCES SOME GREAT EFFECTS. → ↑

↓ CONT. NMOS

BUT WE SHOULD ASK: HOW SMALL CAN WE MAKE THESE DEVICES AND STILL HAVE THEM WORK?

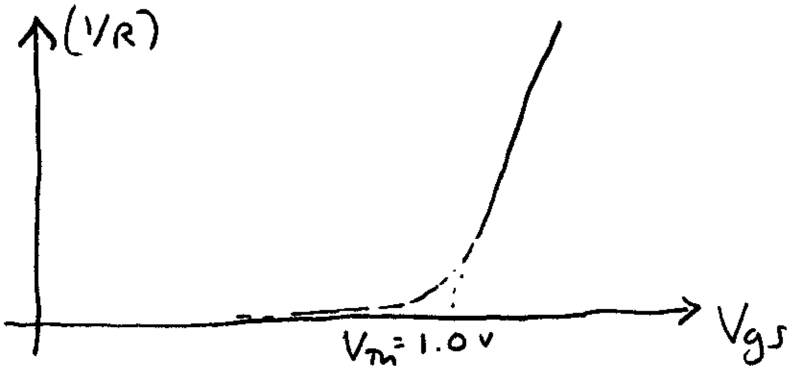
WE MUST SUSPECT THAT THERMAL, STATISTICAL, QUANTUM EFFECTS ARE ULTIMATELY GOING TO MESS THINGS UP!

QUESTION: IF PAT'S FAB WERE NOT LIMITING US, HOW SMALL COULD WE MAKE FET'S AND STILL HAVE THEM WORK?

[MANY FACTORS TO CONSIDER. I'LL DISCUSS SEVERAL OF THE MAJOR ONES, ONE IN SOME DETAIL. IF YOU'RE INTERESTED IN THIS: I SUGGEST READING THE SURVEY PAPER BY KEYES, AND ALSO BROWSING IN CHAPTER 9]

- SUBTHRESHOLD CONDUCTANCE:

IF WE REALLY PLOT DETAILS OF COND. vs V_{gs}, IT IS NOT SIMPLY A STRAIGHT LINE RUNNING DOWN TO V_{TH}, BUT HAS AN EXPONENTIAL TAIL:



Below V_{Th}, the conductance $1/R$ is not zero but depends on V_{gs} and temperature:

$$\boxed{\frac{1}{R} \propto e^{(V_{gs}-V_{Th})/(kT/q)}}$$

T = absolute temp

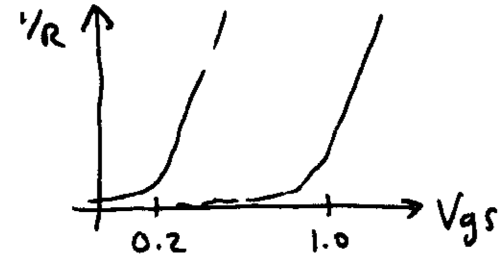
k = Boltzmann's constant

q = charge on electron

At room temperature, $\frac{kT}{q} \sim 0.025$ volts

Thus, at present threshold voltages, an off device, below threshold by perhaps 0.5 volts, is below threshold by $20kT/q$. Thus its conductance is decreased by a factor of $\approx 10^7$ over that when on near threshold. Said another way: if used as a pass-T, Q taking T to pass thru on device will take $10^7 T$ to pass thru off device.

BUT suppose scale down by factor of 5:



[Curves keep ~ same form.
But shifted left.]

Now the OFF FET is DOWN ONLY BY $4kT/Q$, ∴ may have as much as $\frac{1}{100}$ conductance when off as when on.

Use of dynamic storage especially in memories, where stored for many T, will be increasingly harder, espec. below 1 μm. Of course trying to do things statically causes us power dissipation problems. Ah --- you can now see how we're going to get boxed in.

We could scale without continuing to scale voltages say when VDD reaches about 1V. But this also causes us power dissipation problems!

EVEN NOW VOLTAGE SCALING SHOULD BE DONE TO MAX. MOS PERF. SAY TO 3V

Could reduce prob by Reducing T. See interesting paper by Gaenslen, Rideout, Walker CMJ where very small MOSFETs were op. at liquid N temps and measurements confirm improvements.

[But THE TTL comp. kept at 5v

everything as in ratios of #KT

ALL ENERGIES CAN BE SCALED AS $\propto$ KT

Ah, related to low temp operation: Side point:
SO → Reducing temperature also reduces $_{min}$ Esw. Figures quoted were at room temperature. But be careful!

CH9 shows interesting comparison: FET's vs Josephsons:

Won't go into full detail, read if you are interested:

MAYBE DON'T BOTHER WITH THIS

Although switching energy at Device is lowered, you must put in energy into the refrigerator to keep it at the low temp, at least as much as ~~difference~~ resulting from lower T_{wall} operation.

NOW, A COMP DIFF TECH: USE FLUX NOT Q TO ST. INFO.

IBM CS's quote the low Esw of Josephson Junctions at the low temperature environment. It turns out that if you scale FET's down to ~ 1/2 μm, gaining the α^3 improvement in Esw, and operate them at the same temp as J-J's – They will have the same Esw!

But to calculate the energy requirement for a computation, must use T at the heat sink temperature. Refrigerating devices to reduce energy of computation is the logical equivalent of constructing a perpetual motion machine.

So: Viewed as a system: FET system and JJ system will have similar ~~energy~~ switching energy requirements. FET will be simple (operate ≈ room temp.). J-J has advantage of trading the power-delay trade off to lower values of delay (~ 1/50 best FET's). J-Js at quantum limits at ~ 1 μm sizes. Can't be scaled down smaller. ---

So the factor of 100,000 quoted by IBM CS's of switching energy is wiped out by X 1000 (possible) improvement in FET's by scaling, and X 100 due to the perpetual motion machine error.

JUST MENTION THESE BRIEFLY:

OTHER LIMITING FACTORS: (SEE REF BY KEYES, SEE CH9)

- STATISTICAL VARIATIONS IN THRESHOLD VOLTAGE:

AS WE SCALE DOWN, WE'LL FIND THAT $\frac{\Delta V_{Th}}{V_{Th}}$ is proportional to Scaling factor α.

Results from granularity & statistical distribution of substrate impurity charges which determine the threshold voltages.

At same time, # devices increasing. If pullup threshold goes one way, and pulldown another, may end up with inverter which doesn't work. As shown in Ch.9, this may also limit how small supply voltages can be made. In VLSI system contain 10^7 inverters, if we require probability (that all FETs being within threshold limits) = 0.9, may require VDD ≈ 0.7v.

- Quantum Effects. Gate oxide is already only 1000 Å = 0.1μm thick. Positional uncertainty for electrons is related to uncertainty in momentum by

$$\Delta p \Delta x \approx \hbar$$

For energy barrier of ~1eV, calculating corresponding Δp, we find ΔX is about .001 μm. Gate oxides and junction depletion layers must be many times this or electrons will "tunnel" through. Thus we are near a fundamental size limitation due to quantum phenomena.

SUMMARIZING HOW THINGS MAY GO:

	1978	MID-80'S	19XX
MIN FEAT. SIZE (2λ):	6μm	1μm	0.3μm
τ :	0.3 to 1.0 ns	~0.05 to 0.15ns	~0.02 ns to 0.04 ns
E_{SW} :	$\sim 10^{-12}$ J	$\sim 5\times10^{-15}$ J	$\sim 2\times10^{-16}$ J
LOCAL SYNCH SYS CLK. PERIOD (~100τ) :	~30 to 100 ns	~5 to 15 ns	~ 2 to 4 ns

> The mid 80's column we'll probably reach without major hassles. Voltage will be scaled to 1½ or 1V, and power density won't be too much of a problem.

> Subthreshold current will be emerging as a problem, but not within our digital processing structures where "refresh" occurs every 50τ or so.

> Current density will be a rapidly emerging problem, but will be handled with more area devoted to power lines and higher-aspect ratio wires.

- Getting the last order of magnitude out of the technology before fundamental physical limits are finally hit will, however, be a major hassle. It will require close collaboration of researchers spanning the range from CS, to Arch., to E.E., to Device phys., to materials, in order to provide the overall context to help narrow down & select the alternatives to explore.

ON ORDER OF OF COURSE, WE ARE STILL LEFT WITH THE PROBLEM:

WAVELENGTH OF LIGHT! .4 to .7μm MV ≈ .3μm

HOW DO WE MAKE SYSTEMS THIS SMALL?