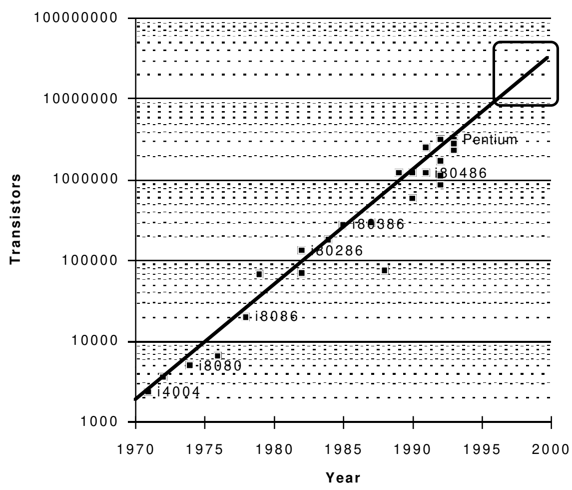
Lecture 34: Portable Systems—Technology Background

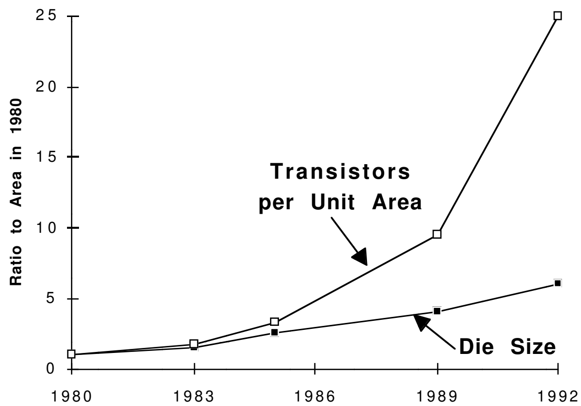
Professor Randy H. Katz
Computer Science 252
Fall 1995

Technology Trends: Microprocessor Capacity



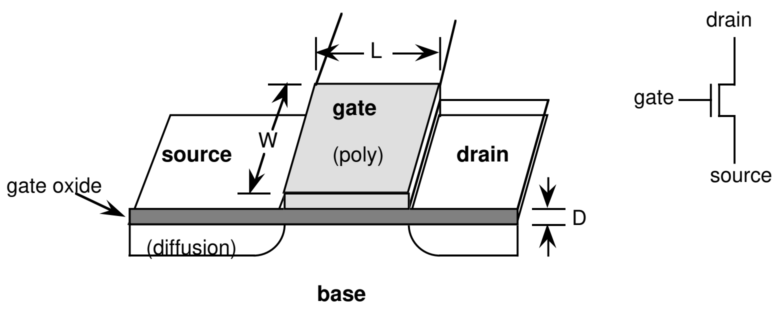
Technology Trends

	Capacity	Speed
Logic	2x in 3 years	2x in 3 years
dRAM	4x in 3 years	1.4x in 10 years
disk	2x in 3 years	1.4x in 10 years



RHK.F95 3

nMOS Transistor

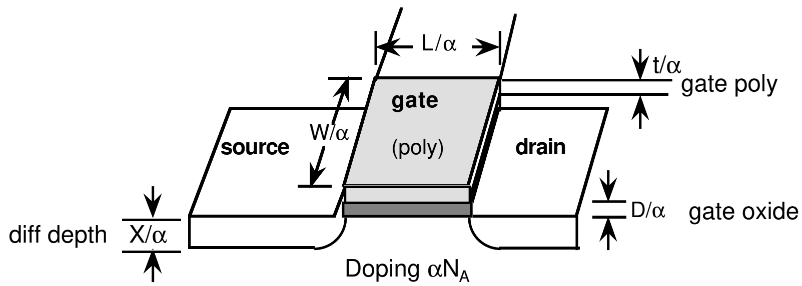


- **Charge on the gate controls the movement of negative charge from source to drain (Vd > Vs) through the *channel***
  - **No charge on gate => open switch**
  - **$V_{gs} > V_{th}$ => conducting path**
- **Size of minimum transistor is determined by minimum width of polysilicon line, minimum feature size, λ**

RHK.F95 4

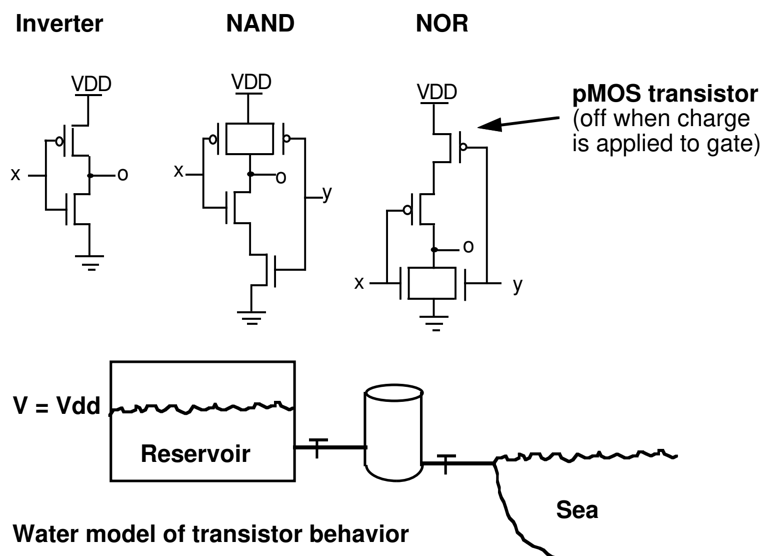
MOS Scaling

- **Scaling factor** α:



RHK.F95 5

CMOS Logic Gates



Water model of transistor behavior

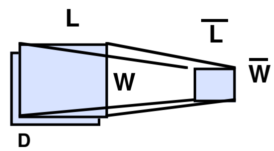
RHK.F95 6

Ideal Scaling

$$\tau \approx \frac{L^2}{V} \qquad C \approx \frac{LW}{D} \qquad I \approx \frac{WV^2}{LD} \qquad P_{sw} \approx \frac{CV^2}{\tau} \qquad P_{dc} \approx IV$$

$$\bar{L} = \frac{L}{\alpha} \quad \bar{W} = \frac{W}{\alpha} \quad \bar{D} = \frac{D}{\alpha} \quad \bar{V} = \frac{V}{\alpha}$$

(and increase doping)



$$\frac{\bar{\tau}}{\tau} = \frac{\left(\frac{\left(\frac{L}{\alpha}\right)^2}{\left(\frac{V}{\alpha}\right)}\right)}{\left(\frac{L^2}{V}\right)} = \frac{1}{\alpha} \qquad \frac{\bar{C}}{C} = \frac{\left(\frac{\frac{L}{\alpha}\frac{W}{\alpha}}{\left(\frac{D}{\alpha}\right)}\right)}{\left(\frac{LW}{D}\right)} = \frac{1}{\alpha} \qquad \frac{\bar{I}}{I} = \frac{\left(\frac{1}{a^3}\right)}{\left(\frac{1}{\alpha^2}\right)} = \frac{1}{\alpha} \qquad \frac{\bar{P}}{P} = \frac{1}{\alpha^2} \qquad \frac{\bar{N}}{N} = \alpha^2$$

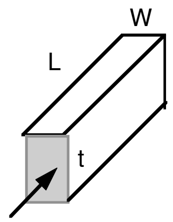
faster **less capacitance** **less current** **less power** **more gates**

power density? **What if you don't scale the voltage?**

MOS Scaling Summary

Parameters		**Scaling**	**Effects**
Length	L	$1/\alpha$	Channel length to width ratio unchanged
Width	W	$1/\alpha$	Gate area … reduced by $1/\alpha^2$
Gate Oxide	D	$1/\alpha$	Gate dielectric thickness reduced by $1/\alpha$
Junction Depth	Xj	$1/\alpha$	Gate cap sq Cg reduced by $1/\alpha$
Layer Thickness	t	$1/\alpha$	Parasitic capacitances (fXj) reduced by $1/\alpha$
			Resistance layer thickness reduced by $1/\alpha$
Subs. Doping	N	α	Resistivity ρ reduced by $1/\alpha$
			Sheet resistance Rs = ρ/t unchanged
			Time delay τ = Rs x sq Cg reduced by $1/\alpha$
			Inverter/gate delay reduced by $1/\alpha$
Supply Voltage	Vdd	$1/\alpha$	Current reduced by $1/\alpha$
			X-section of conductors reduced by $1/\alpha^2$
			Current density increased by α
			Logic levels reduced by $1/\alpha$
			Power diss Pd reduced by $1/\alpha^2$
			Power-speed product reduced by $1/\alpha^3$
			Switching Energy/circuit fCgVdd^2 reduced by $1/\alpha^3$
			Components/unit area increased by α^2
			Complexity/chip increased by α^2
			Power diss/unit area unchanged

Wire and Interconnect Scaling



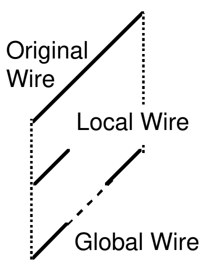
Wire of length L, width W, thickness t, and resistivity ρ

Resistance unscaled: $R_u = \rho L / (W t)$

Resistance scaled: $R_{sc} = \dfrac{\rho L / \alpha}{W t / \alpha^2}$

$R_{sc} = \alpha R_u$

Sheet Resistance $R_s = \rho / t$



$R = R_s L / W$: R_s must increase by α when t is scaled down by $1/\alpha$

I down by $1/\alpha$, IR drops remain constant
Since voltage scales down by α, this scales up *relatively* by α

$\tau_{sc} = (\alpha R)(C/\alpha) = R C = \text{constant} = \tau$
Time delays of local wires do not scale!
Time delays of global wires scale UP by α
Local communication is at a premium

RHK.F95 9

Wire and Interconnect Scaling Summary

Current density: $J = I / (W t)$

Scaled current density: $J_{sc} = \dfrac{(I / \alpha)}{(W t / \alpha^2)} = \alpha J$

Parameters	**Scaling Factor**
Line Resistance R	α
Line Voltage Drop Vd	1
Normalized Line Volt drop Vd/V	α
Current Density J	α
Normalized contact, voltage drop Vc/V	α^2

RHK.F95 10

Transistor Timing Model

$$I_{ds} = \frac{\text{charge in transit}}{\text{transit time}} = \frac{Q}{\tau} = \frac{-C_g\,(V_{gs} - V_{th})}{\tau}$$

τ = time for charge to travel across channel

$$\tau = \frac{L}{\text{velocity}} = \frac{L}{\mu E} = \frac{L^2}{\mu V_{ds}} \qquad C_g = \frac{\varepsilon A}{D} = \frac{\varepsilon W L}{D}$$

mobility **(cm^2 / volt-sec)** ***permittivity***

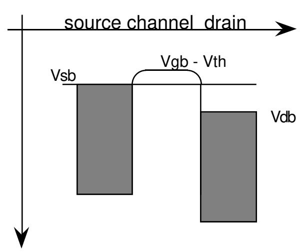
$$I_{ds} = \frac{\mu e W}{LD}\,(V_{gs} - V_{th})\,V_{ds}$$ **in resistive region**

$$I_{ds} = \frac{\mu e W}{LD}\,(V_{gs} - V_{th})^2$$ **in saturation**

RHK.F95 11

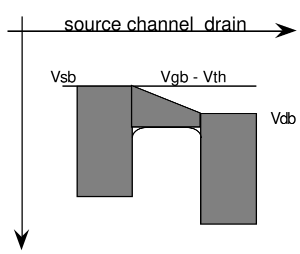
Model of Behavior

Interface potential



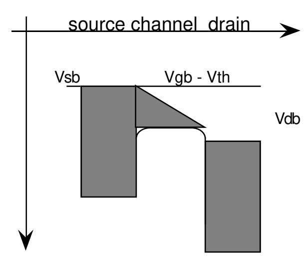
Cut Off: $V_{gs} < V_{th}$ ($V_{gb} - V_{th} < V_{sb}$)

$I_{ds} = 0$



Resistive: $V_{gs} \geq V_{th}$, $V_{ds} < V_{gs} - V_{th}$

$I_{ds} \propto V_{ds}$



Saturation: $V_{gs} \geq V_{th}$, $V_{ds} \geq V_{gs} - V_{th}$

$I_{ds} \propto (V_{gs} - V_{th})$

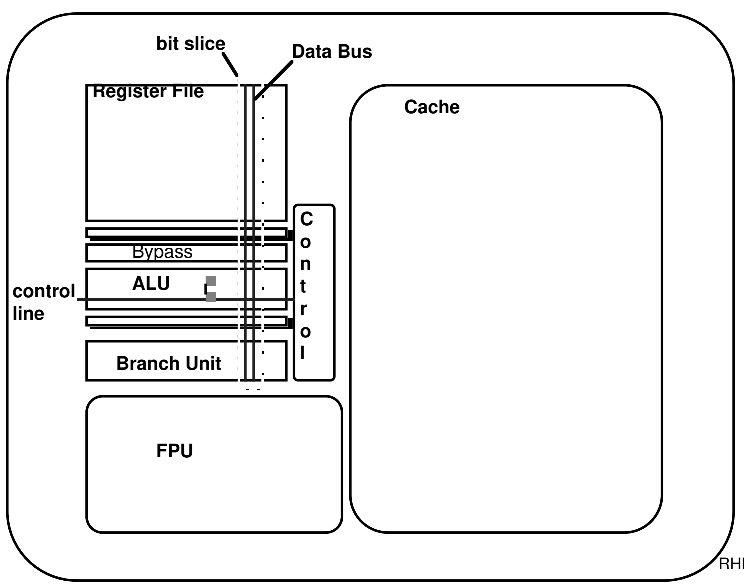
RHK.F95 12

Performance Capacity

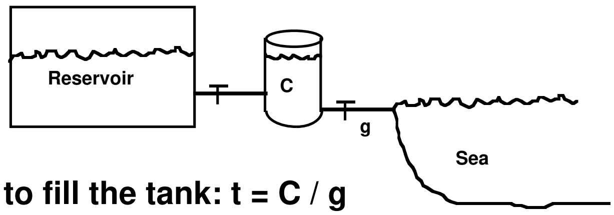
- Feature size shrinks about 10% per year
  - Switch speeds improve by 1.2x per year
  - Switching density increases by 1.2x per year
- Die area increases by about 20% per year
  - Total computing power increases by 1.73 x per year (1.2^3)

If we can utilize every gate all the time!

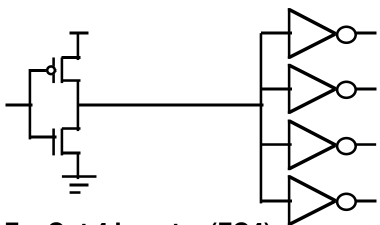
Processor Structure



Fanout



Time to fill the tank: t = C / g
- C: capacity of the tank (capacitance)
- g: pipe flow (transistor strength)



FanOut 4 Inverter (FO4)

FO4 Timing Model

Gates:

Inv	1.0	Passgate	0.5 (FO2)
Nand2	1.5	Mux2	1.5
Nand4	2.0	Mux4	2.0
Nor2	1.5		
Nor4	3.0	TriBuff	2.0
Latch	1.5		2.5 (FO16)

Larger Units

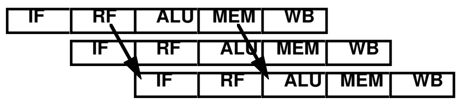
Adder (32 bit, not ALU)	12 - 14
RegFile	11 - 13
Precharge compare	1.25 / 4 bits
Memory (address to data, 8Kbyte)	16-20

Wires (32 bit data path in .8 μ, 1mm x 4mm)

Control wire (1 mm X 2) = 10 Cg + 32 gates > 3 FO4

Global data bus = 20 Cg + GateCap + DiffCap

Typical Critical Paths

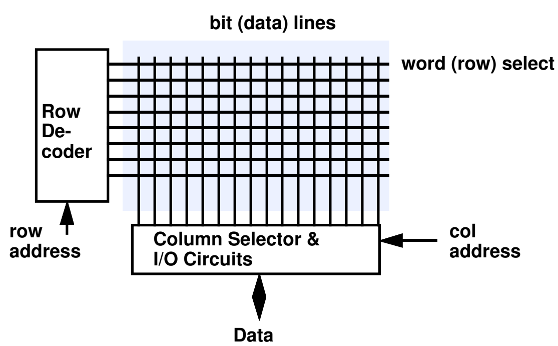


- Sparc / MIPS R3000 style pipelines ~ 24-25 FO4
- Alpha, R4400 style ~ 20 FO4

- Where is the critical path?
  - Register file access
  - ALU
  - Address -> cache
  - Cache data -> DP
  - Branch condition and address
  - TLB

RHK.F95 17

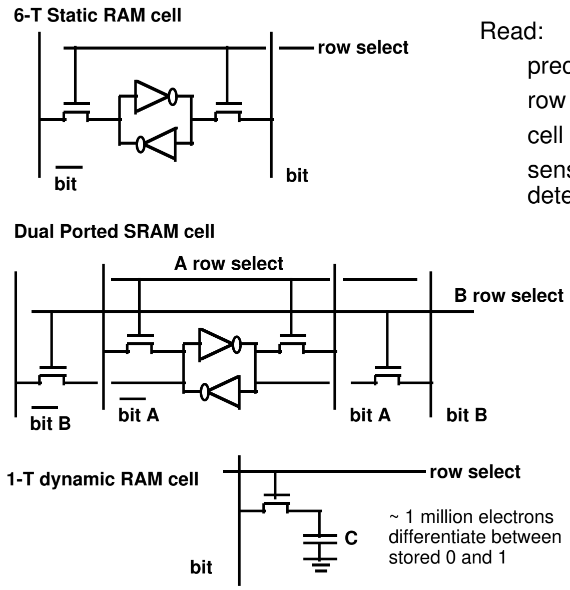
What goes in a RAM?



Spliting the address into row x col allows aspect ratio and speed to be controlled.

RHK.F95 18

RAM Cells



Read:

precharge bit and $\overline{\text{bit}}$ to Vdd

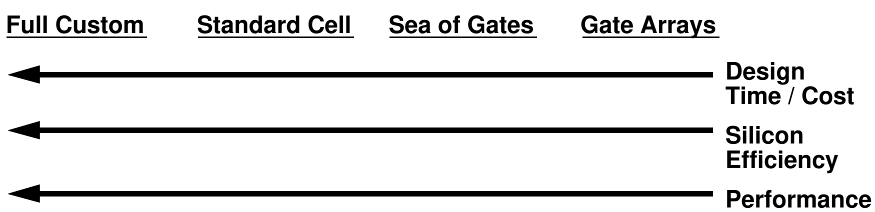
row select

cell pulls one line low

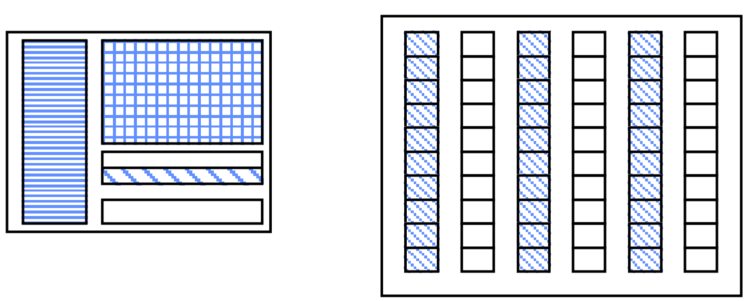
sense amp on each column detects differential signal

RHK.F95 19

VLSI Design Styles



NRE, Inventory Risks, Design Risks, Turn-around



RHK.F95 20