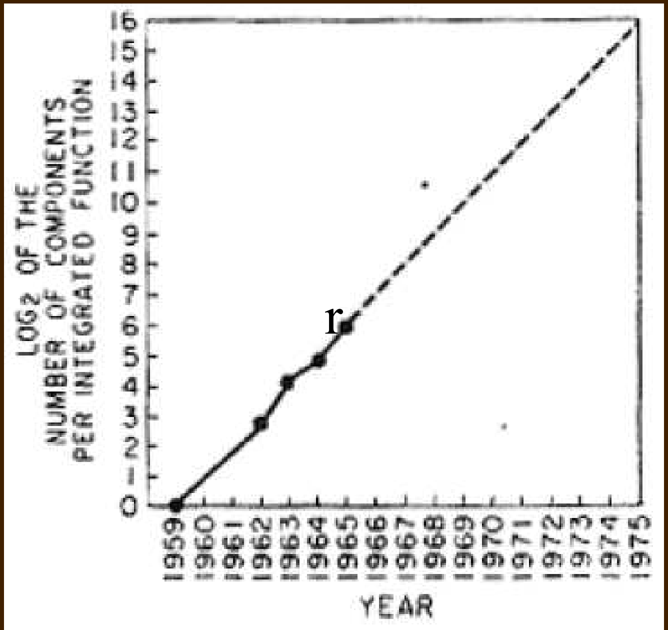
The GHz Race Is Over

Mark Greenstreet, CpSc 421, Term 1, 2006/07

- Moore's Law
- Power is the Problem
- The Future is Parallel
- Other Technologies

Moore’s Law

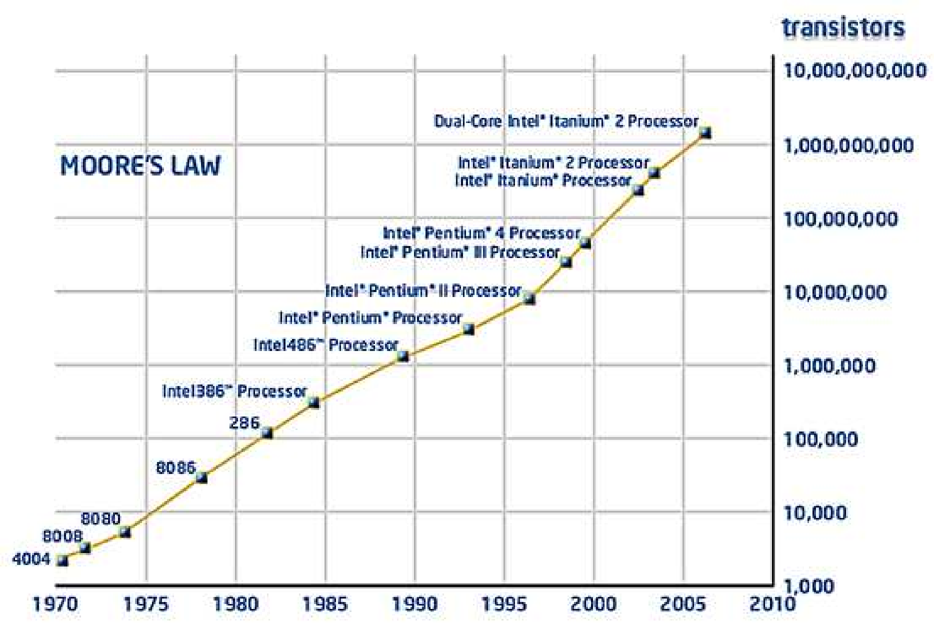


- In 1965, Gordon Moore wrote a now famous paper:

  Cramming More Components Onto Integrated Circuits

- Integrated circuits were very new, and very small by today’s standards. Moore projected that integrated circuits would double in transistor count every year for the next decade.

Moore’s Law in 2006



- Moore’s prediction has been surprisingly accurate:
  - In 1965, chips had $64 = 2^6$ transistors.
  - Today, chips have up to about 1 billion (2^{30}) transistors.
  - That’s a doubling time of $31\text{years}/(30-6) \approx 1.3\text{years}$.

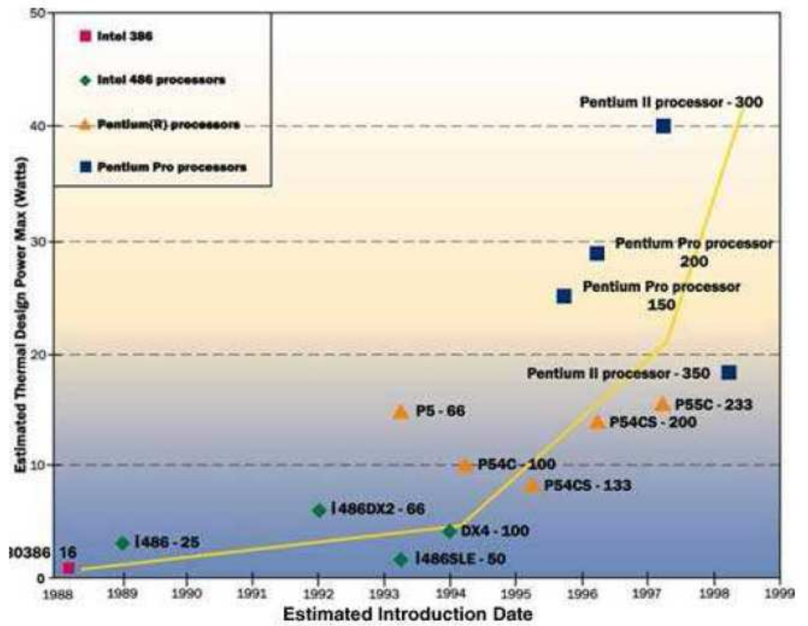
Moore’s Law in 2006

- Moore’s law is largely economic:
  - In 1965, there was lots of physical opportunity to make smaller transistors, and thus more per chip.
  - Making chips with more transistors → more functionality → more products → more revenue → the money to build fabrication plants that produce chips with more transistors.
  - It’s a positive feedback cycle – that’s the cause of the exponential growth.

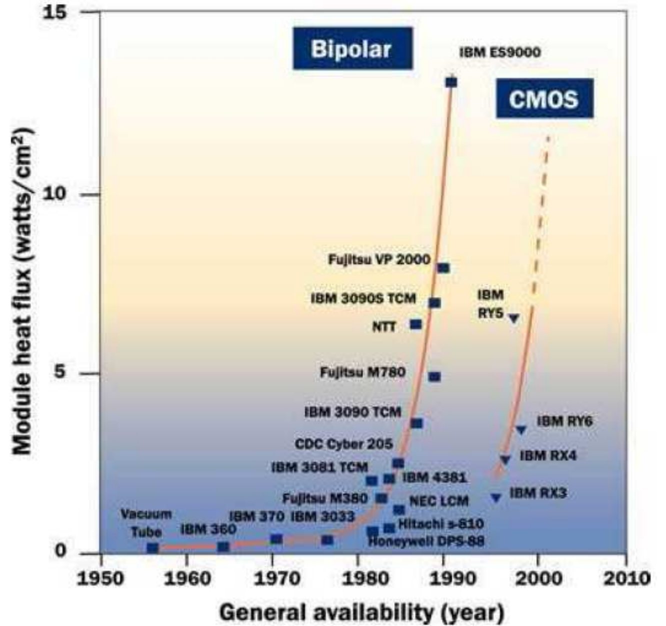
No Exponential Is Forever

- $\log(\text{the size of the universe})$ isn't very big, for any reasonable "size" of the universe.
- What are the limits to Moore's Law?
  - Economics: the cost of fabrication plants and mask sets has been increasing exponentially as well (with a somewhat smaller exponent).
  - Atoms: The first chip designs I worked with (1980) had a 4.5μ gate length. Now, chips are being manufactured with 65nm gates, and fabricated with 45nm gates. That's a factor of 70–100 shrink. Another factor of 70 would be 1nm (10Å) gates – about 4 silicon atoms.
  - Power: Issues from computer architecture, thermodynamics, and quantum mechanics all push power up. There are practical/economical limits to how fast the heat can be removed.

Power Is Today’s Problem

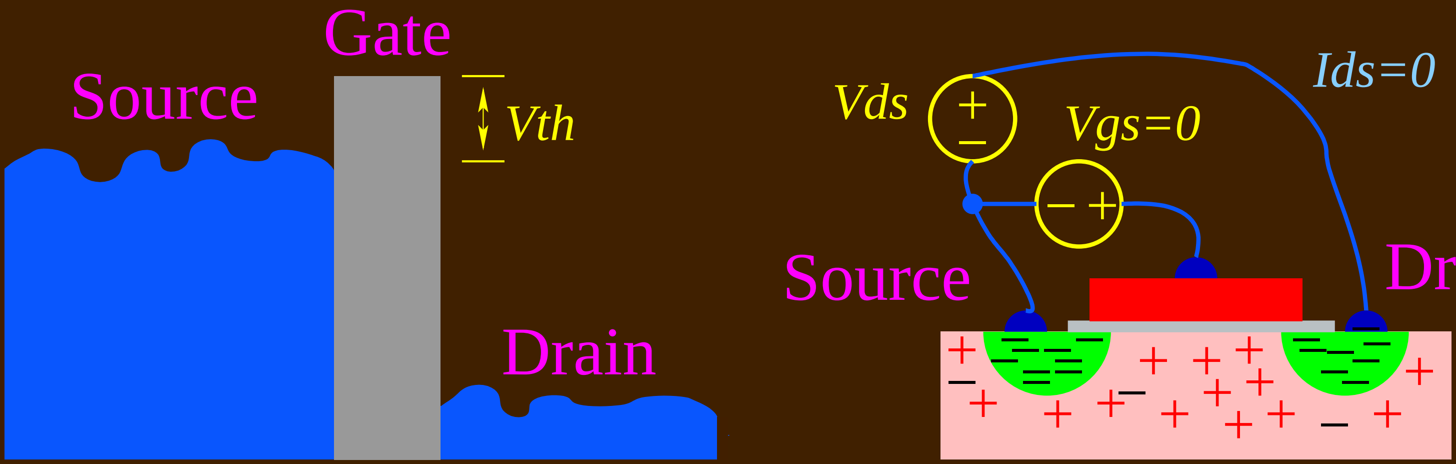


Intel Processors



IBM Processors

Dam Transistors

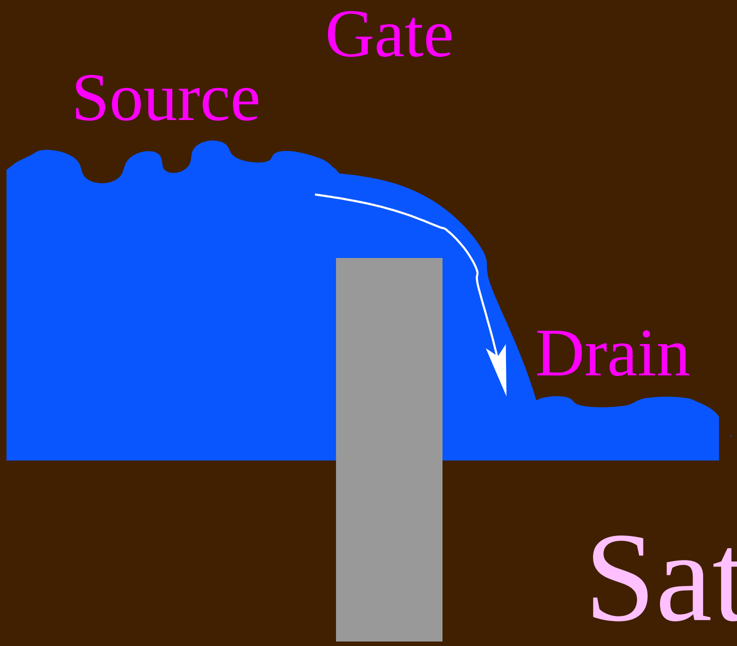


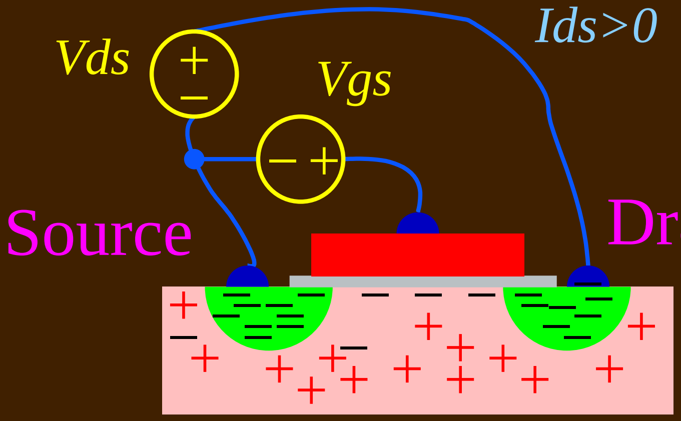
Cut−off

Cut-Off: The dam is higher than the upper resevoir:

$$V_{gs} - V_{th} \leq 0, \quad I_{ds} = 0$$

Dam Transistors



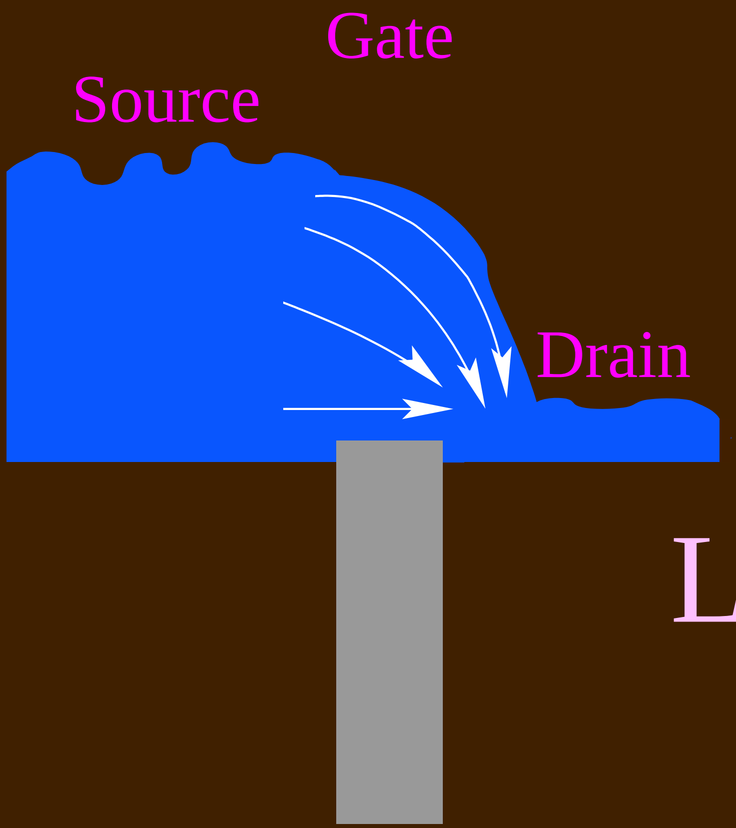


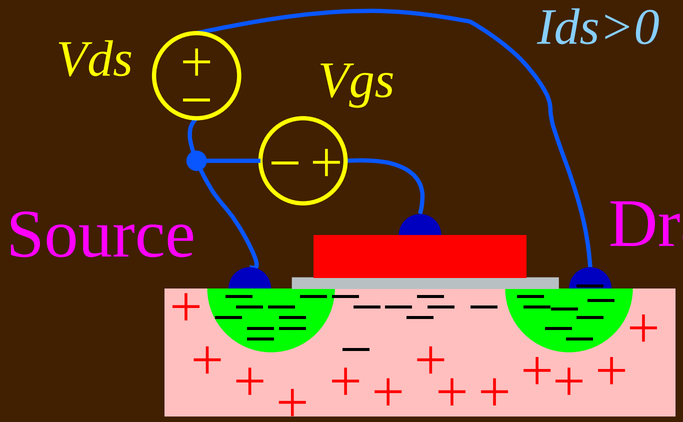
Saturation

Saturation: The dam is between the two resevoirs:

$$0 < V_{gs} - V_{th} \leq V_{ds}, \quad I_{ds} = \frac{K}{2}(V_{gs} - V_{th})^2$$

Dam Transistors





Linear

Linear: The dam is lower than the bottom resevoir:

$$V_{ds} < V_{gs} - V_{th}, \quad I_{ds} = K * (V_{ge} - \frac{1}{2}V_{ds})V_{ds}$$

Physics Review

- Force, Energy and Power:
  - Force: How hard you push on something. Measured in Newtons. $1\text{Newton} = 1\text{kgm}/\text{sec}^2$, $f = ma$.
  - Energy: Force times distance. Measured in Joules. $1\text{Joule} = 1\text{Newton} * 1\text{meter}$, $E = f * d$.
    Note that simple machines (levers, pulleys, bicycle gears, etc.) preserve this product. You expend the same energy lifting 1 kilogram 2 meters as lifting 2 kilograms 1 meter.
  - Power: Energy/time. Measured in Watts. $1\text{Watt} = 1\text{Joule}/\text{secondmeter}$, $P = E/t$.
    A 1 Watt motor can lift 1 kilogram 1 meter in 9.8 seconds.
    A 10 Watt motor can lift 1 kilogram 1 meter in 0.98 seconds.
    The factor of 9.8 is the acceleration of the earth's gravity.
- Volume, Pressure and Energy:
  - Volume: how much space something occupies. Measured in meters^3.
  - Pressure: force/area. Measured in Pascals. $1\text{Pascal} = 1\text{Newton}/\text{meter}^2$,
  - Energy $=$ Volume $*$ Pressure.

Electricity

- Charge: the quantity of electrons, measured in Coulombs.
  -1 Coulomb $= 6.24 * 10^{18}$ electrons. Analagous to volume.
- Voltage: a force applied to charges, meaasured in Volts.
  1 Volt $=$ 1 Joule/Coulomb. Analgous to pressure.
  Note that pressure*volume = energy, and voltage*charge = energy.
- Current: rate of flow of charge, measured in Amperes.
  1 Ampere $=$ 1 Coulomb/Second. Analgous to flow of water (e.g. liters/second).
- Resistance: hindering the flow of electricity, measured in Ohms (Ω). 1 Ohm $=$ 1 Volt/Ampere.
- Capacitance: the ability to store charge. Like filling up a water tank from the bottom: the more water that is in the tank, the harder you have to pump to put more water in. Measured in Farads.
  1 Farad $=$ 1 Coulomb/Volt.

Formulas for Electricity

Ohm's Law:	$V = I * R$	
Time constants:	$\tau = R * C,$	τ is time for capacitor to reach
Charge in C:	$Q = C * V$	
Energy in C:	$E = \frac{1}{2} C * V^2,$	by integration
Power in C:	$P = \frac{1}{2} * C * V^2 * f,$	

where $C =$ capacitance; $E =$ energy; $f =$ frequency; $I =$ current; $Q =$ charge; $R =$ resistance.

Simple Scaling

- Shrink every thing by a factor of α:

What	Scaled Value	Example: $\alpha = 2$
Linear Dimension	$\frac{1}{\alpha}$	$\frac{1}{2}$
logic gates/cm^2	α^2	4
voltage	$\frac{1}{\alpha}$	$\frac{1}{2}$
capacitance (per logic gate)	$\frac{1}{\alpha}$	$\frac{1}{2}$
resistance (transistor)	1	1
frequency ($\propto 1/(RC)$)	α	α
energy (per logic gate, $\propto CV^2$)	$\frac{1}{\alpha^3}$	$\frac{1}{8}$
power (per logic gate, $\propto CV^2 f$)	$\frac{1}{\alpha^2}$	$\frac{1}{4}$
power (total)	1	1

- Everything should get better as transistors get smaller ☺.

What Went Wrong?

- Voltage hasn't dropped by linear scaling model:
  - Operating voltage was 5 volts for the 4.5μ design I worked on in 1980 until $0.35\mu = 350$nm designs in about 1995. Higher voltages allow higher speed (but more power).
  - Voltages aren't dropping for transistors smaller than 90nm because of leakage currents.
  - Thus, a 65nm process operates a 1V, when simple scaling (from 4.5μ) predicts 0.06V, a factor of 16 over prediction. Power goes as V^3, about a factor of 4000.
- Frequency has gone up faster than linear scaling, even when accounting for voltage cheating:
  - Better machine architecture.
  - Marketing.
- More interconnect layers means more capacitance an more power.
  - 1 metal layer in 4.5μ.
  - 11–13 metal layers typical in 65nm.

The extra interconnect is needed to connect together the huge number of logic gates.

Voltage Scaling

- What happens if we change V?
  - f is roughly propotional to V.
  - P is roughly propotional to $V^2 f$.
  - Divide V by α and power goes down by α^3.
- The potential of parallelism
  - Divide one task into two equal pieces.
  - Run the two pieces on processors running at half the voltage and half the speed but one-eighth the power.
  - One quarter the total power.
  - Or, run each processor at half the power, $\sqrt[3]{1/2} \approx 0.79$ the voltage, and thus 79% of the speed. Get the job done 1.6 times faster for the same power.

The Potential of Parallelism

- Parallelism is the only path we have currently to improved computer performance.
- This is why AMD and Intel are celling dual and quad core processors.
- SUN and IBM have pushed this even further with Niagra and Cell.
- But, parallelism is hard to exploit:
  - Reasonably well-understood for a few problems: database servers, web-servers, scientific computing.
  - Pretty good use in computer games – lots of simulation and graphics rendering, i.e. numerical computing.
  - A big challenge for applications programming.

Consequences

- Parallelism is in your future. Applications that can exploit parallelism will be the biggest area of innovation in the next 10 years.
- Impact on all of computer technology:
  - HCI will become even more important. If the next version can't have more features (because processor speed and memory sizes are at their limits), usability will become the key aspect for gaining market share.
  - Devices, circuits, and micro-architecture will matter as we try to squeeze what we can out of silicon and transition to other technologies.
  - Software engineering, operating systems, compilers and programming languages will all evolve to reflect the critical role of parallelism.

CpSc 418 Preview

- Parallelism in CPUs: pipelined and superscalar architectures.
- Parallel machines: message passing, shared memory.
- Transistors, technology scaling, and power.
- Power aware architectures: chip-multiprocessors, processor arrays, etc.
- Energy and algorithms.
- Future technology directions: carbon nanotubes, super-cooled computers, quantum computers, biological computing, etc.
- With each of these, we'll try to look at the likely impact on computer science, programming, and applications.