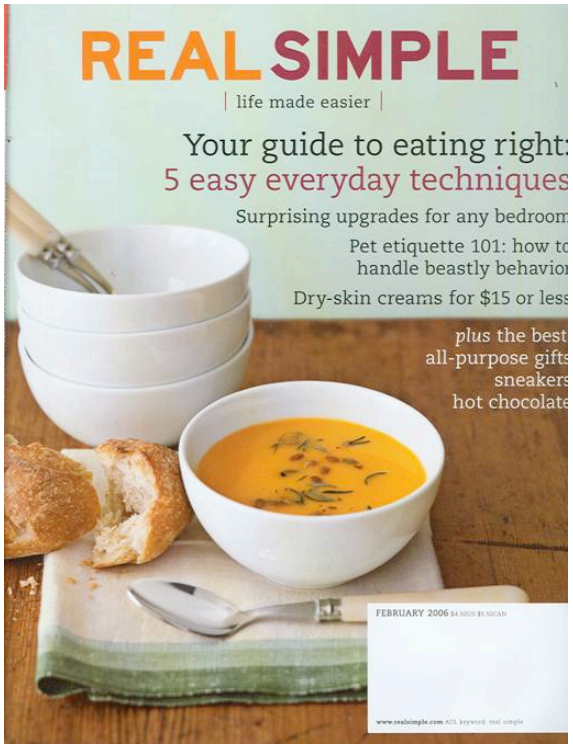
Lecture 01
Introduction to CSE 30321

Huh?

All of the following are magazines that are regularly delivered to the Niemier household.



You can learn about good routes to run if you're visiting Chicago...

You can learn about the best ice cream...

road test | ICE CREAM POPS

joy sticks

Never send an adult to do a child's job. *Real Simple* gathered a group of junior gourmands, ages five through 12 (plus a handful of their parents), and asked them to rate ice cream pops on a scale of "gross" to "awesome." After the feeding frenzy, four frosty wonders remained. Here's what scored highest on the taste-o-meter.

best take on a classic

HÄAGEN-DAZS VANILLA & MILK CHOCOLATE BAR

Hailed by one tester for its "rich, thick, and buttery" filling, this winner got points for how the "nice, chunky chocolate coating crackled in my mouth and melted into the vanilla."

TO BUY: $4.20 for three 3.67-ounce bars.

most decadent

ESKIMO PIE VANILLA NESTLÉ CRUNCH BAR

"Crunchy and yumful" is how one grade-schooler rated this treat. "The thin coating gets me to the ice cream fast." Adults praised the pop's "pleasing but not too rich" interior.

TO BUY: $4 for six 3-ounce bars.

best citrus blast

POPSICLE CREAMSICLE

"Very refreshing," praised a panelist of this old-school icon. Said an eight-year-old: "Tastes like frozen orange juice, but smooth because of the vanilla ice cream."

TO BUY: $3.30 for 12 1.65-ounce bars in two flavors.

most virtuous

THE SKINNY COW MINI FUDGE BAR

From the intense flavor—"it tastes super chocolaty," raved an aficionado—you'd never guess this pop packs only 50 calories. "Icy and rich" was the general consensus.

TO BUY: $5 for 12 1.5-ounce bars.

how to stop brain freeze

Talk about a buzz kill: You're enjoying a delicious Slurpee or margarita when you get a piercing headache. This probably happens when capillaries in the roof of your mouth become dilated from coming into contact with an icy substance, says Mark W. Green, M.D., director of headache medicine at the Columbia University College of Dental Medicine. A message is then sent to your brain, and in response blood vessels can swell, resulting in a brain freeze. The cure? Simply press your tongue to the roof of your mouth to warm it, or drink a glass of warm water.

click **Top scoops** For the best sorbets, go to www.realsimple.com/sorbet.

REAL FEB / AUG 07 2006 132

You can read about Pat Robertson and Al Sharpton advocating ways to reduce the effects of global climate change!

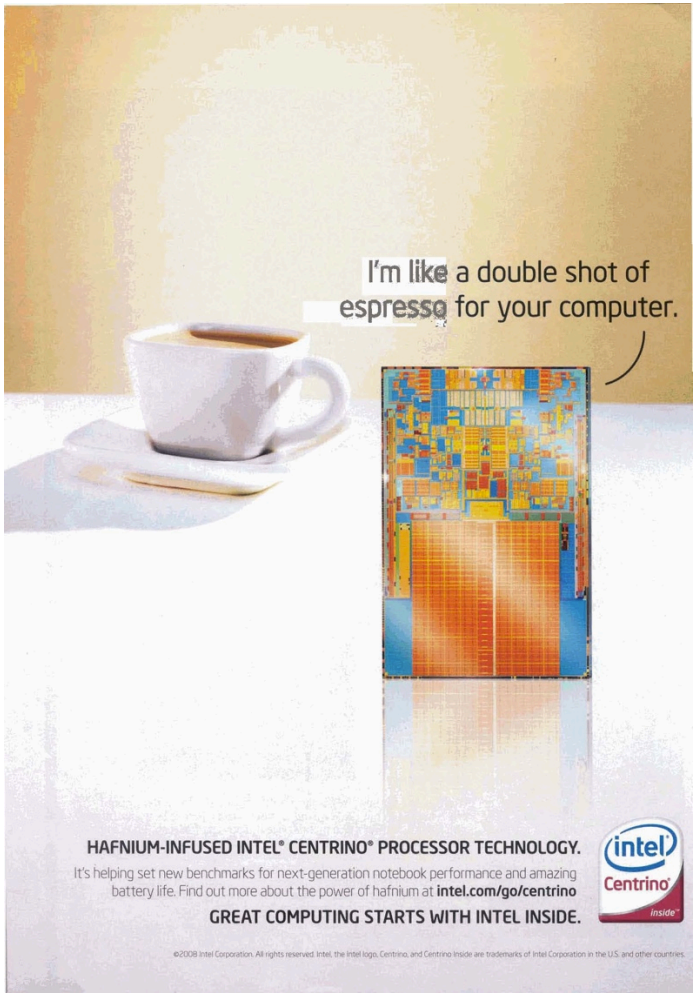
Rev. Al Sharpton and Rev. Pat Robertson, Virginia Beach, VA

It's American to disagree. It's also American to come together in the face of a challenge. And few challenges are as urgent as global climate change. Take one minute to join us at wecansolveit.org and add your voice to millions of others. Together we can solve the climate crisis.

Join we today.

If you're in the market for a wide-body jet, Boeing has just the thing...

And if you're looking to buy a computer, Intel suggests their dual core Centrino chip...



As you might guess, this brings us to CSE 30321...

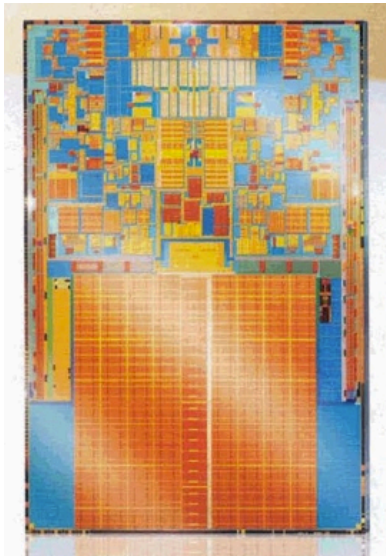
In this ad, you're essentially looking at a picture of one of Intel's computer architectures...

... and computer architecture is what this class is about.

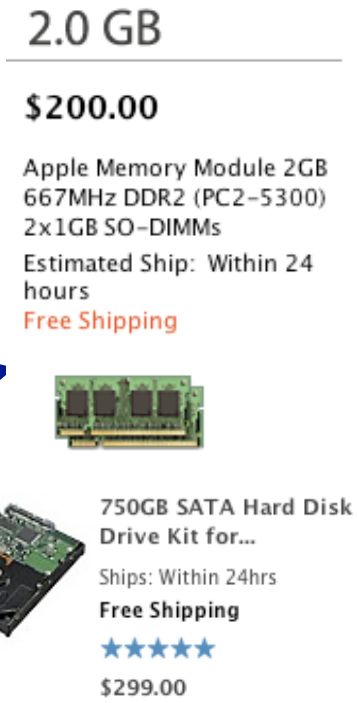
So, what are the goals of this course?

- At the end of the semester, you should be able to...
  - ...describe the fundamental components required in a single core of a modern microprocessor
    - (Also, explain how they interact with each other, with main memory, and with external storage media...)

Example



How do on-chip memory, processor logic, main memory, disk interact?



So, what are the goals of this course?

- **At the end of the semester, you should be able to...**
  - **...compare and contrast different computer architectures to determine which one performs better...**

Example

AMD Smarter Choice

Processor	AMD Athlon™
Model	3200+
OPN Tray	ADA3200AEP5AR
OPN PIB	ADA3200BOX
Operating Mode 32 Bit	Yes
Operating Mode 64 Bit	Yes
Revision	CG
Core Speed (MHz)	2000
Voltages	1.50V
Max Temps (C)	70
Wattage	89 W
L1 Cache Size (KB)	128
L1 Cache Count	1
L2 Cache Size (KB)	1024
L2 Cache Count	1

Intel® Pentium® Dual-Core processor

The Intel® Pentium® dual-core processor delivers great performance, low power enhancements, and multitasking for everyday computing.

» Learn more

Processor Number[1]	Architecture	Cache	Clock Speed	Front Side Bus	Dual-core	Enhanced Intel SpeedStep® Technology[2]	Execute Disable Bit[3]	Intel® 64[4]
E2220	65 nm	1MB L2	2.40 GHz	800 MHz	✓	✓	✓	✓
E2200	65 nm	1MB L2	2.20 GHz	800 MHz	✓	✓	✓	✓
E2180	65 nm	1MB L2	2.00 GHz	800 MHz	✓	✓	✓	✓
E2160	65 nm	1MB L2	1.80 GHz	800 MHz	✓	✓	✓	✓
E2140	65 nm	1MB L2	1.60 GHz	800 MHz	✓	✓	✓	✓
T2330	65 nm	1MB L2	1.60 GHz	533 MHz	✓	✓	✓	✓
T2310	65 nm	1MB L2	1.46 GHz	533 MHz	✓	✓	✓	✓
T2130	65 nm	1MB L2	1.86 GHz	533 MHz	✓	✓	✓	
T2080	65 nm	1MB L2	1.73 GHz	533 MHz	✓	✓	✓	
T2060	65 nm	1MB L2	1.60 GHz	533 MHz	✓	✓	✓	
T2370	65 nm	1MB L2	1.73 GHz	533 MHz	✓	✓	✓	✓

If you want to do X, which processor is best?

So, what are the goals of this course?

- At the end of the semester, you should be able to...
  - ...design a processor architecture to meet a specific performance target...

Example

Find by Feature

Processor	AMD Athlon™ X2 Dual-Core
Model Number	6400+
Frequency (MHz)	3200
L2 Cache Size (KB)	1024
Socket	AM2
Stepping	F3
Manufacturing Tech (CMOS)	90nm SOI
Wattage (W)	125 W
System Bus (MHz)	2000
AMD Business Class	No

Find by Feature

Processor	AMD Athlon™ X2 Dual-Core
Model Number	5600+
Frequency (MHz)	2900
L2 Cache Size (KB)	512
Socket	AM2
Stepping	G2
Manufacturing Tech (CMOS)	65nm SOI
Wattage (W)	65 W
System Bus (MHz)	2000
AMD Business Class	No

You might choose to add more or less on-chip memory...

So, what are the goals of this course?

- **At the end of the semester, you should be able to...**
  - **...understand how code written in a high-level language (e.g. C) is eventually executed on-chip...**

Example

In C:

```
void insertionSort(int numbers[], int array_size)
{
  int i, j, index;

  for (i=1; i < array_size; i++)
  {
    index = numbers[i];
    j = i;
    while ((j > 0) && (numbers[j-1] > index))
    {
      numbers[j] = numbers[j-1];
      j = j - 1;
    }
    numbers[j] = index;
  }
}
```

In Java:

```
public static void insertionSort(int[] list, int length) {
    int firstOutOfOrder, location, temp;

    for(firstOutOfOrder = 1; firstOutOfOrder < length; firstOutOfOrder++) {
        if(list[firstOutOfOrder] < list[firstOutOfOrder - 1]) {
            temp = list[firstOutOfOrder];
            location = firstOutOfOrder;

            do {
                list[location] = list[location-1];
                location--;
            }
            while (location > 0 && list[location-1] > temp);

            list[location] = temp;
        }
    }
}
```

Both programs could be run on the same processor... how does this happen?

A tangent...

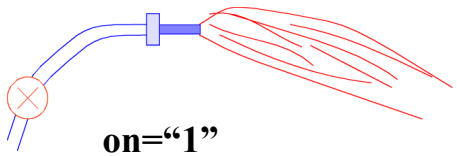
- We'll talk about 2 more course goals in a little bit, but right now, I'd like to ask the class a few questions...
  - Question 1:
    - How many people are EE, CPEG, CS, other?
  - Question 2:
    - By major, does anyone have any definitive thoughts about what they want to do after graduation?
  - Question 3:
    - Preface: last slide talked about SW...
    - How many people are more interested in the SW side of CSE than the HW side of CSE?
  - Question 4:
    - How many people view CSE 30321 as more a "HW course"
    - How many people think other more "SW oriented" courses are more relevant for their major?

Let's digress...

- I asked the questions on the last slide not just to gauge interest, but to bring up an important point...
  - For last 20 years, if interested in SW, computer architecture was probably *not* the most important class for you.
- But...changes in technology are having a profound impact on conventional/established computer architectures
  - We're presently at the very beginning of this storm...
- We'll need significant engagement from programmers to continue the processor performance scaling trends of the last 40 years...
  - ... this will impact your career...

A little history... Zuse's paradigm

- Konrad Zuse (1938) Z3 machine
  - Use binary numbers to encode information
  - Represent binary digits as on/off state of a current switch



on="1"

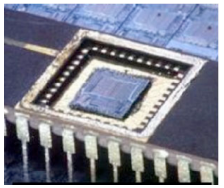
The flow through one switch turns another on or off.

Exponential down-scaling

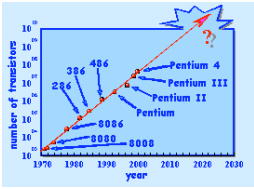
Electromechanical relay

Vacuum tubes

Solid-state transistors



CMOS IC



Result: exponential transistor density increase…

A little history... programs

- **Stored program model has been around for a long time...**

First Draft of a Report
on the EDVAC

by

John von Neumann

Contract No. W–670–ORD–4926

Between the

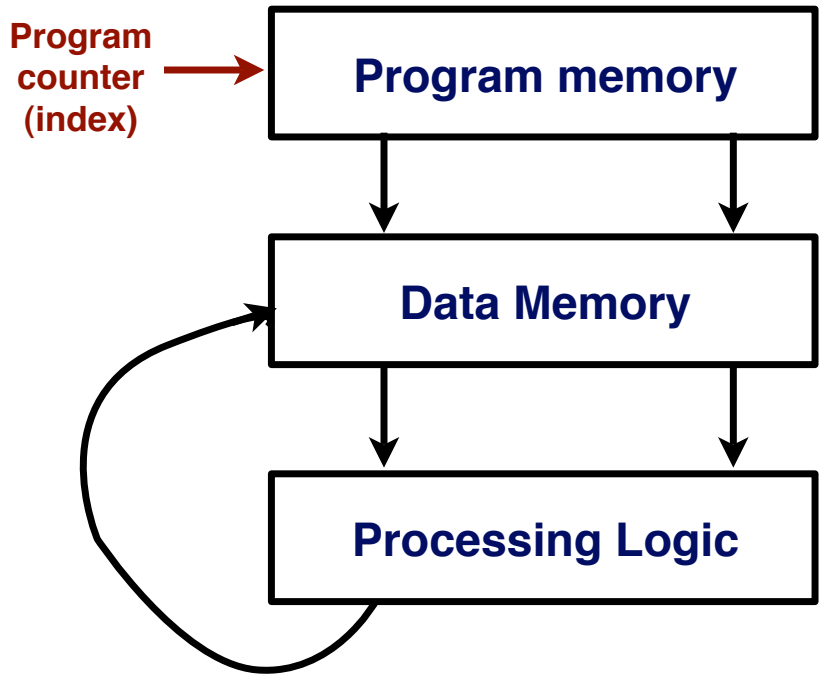
United States Army Ordnance Department

and the

University of Pennsylvania

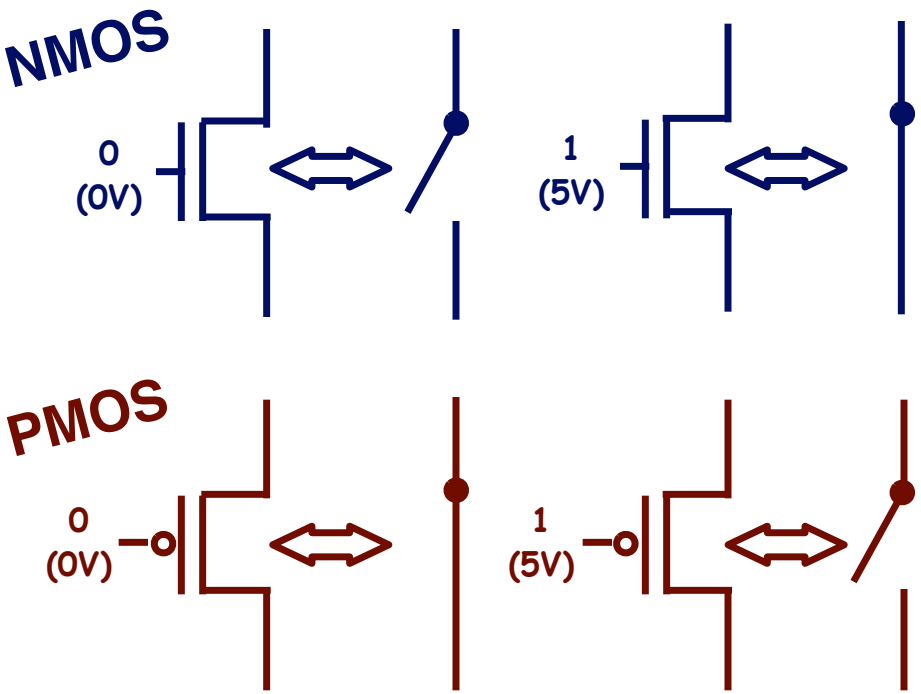
Moore School of Electrical Engineering
University of Pennsylvania

June 30, 1945

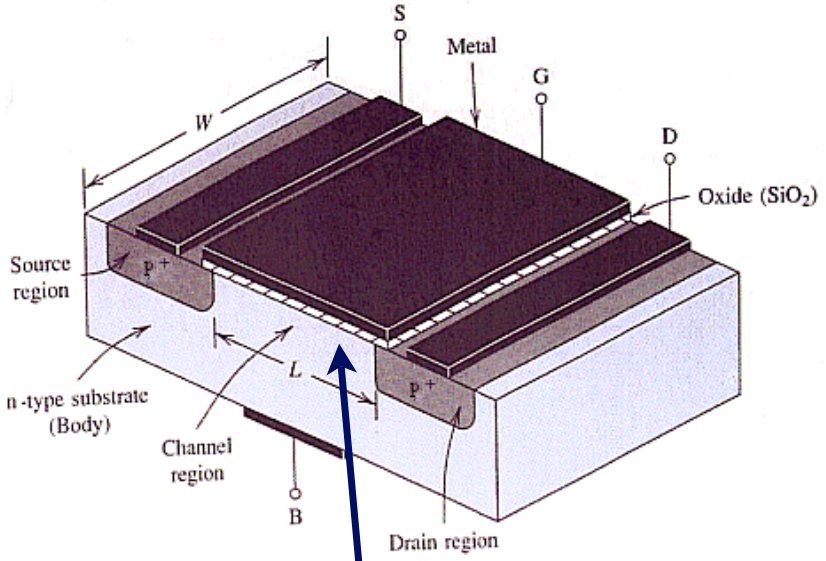


Transistors used to manipulate/store 1s & 0s

Switch-level representation



Cross-sectional view



(can act as a capacitor storing charge)

Using above diagrams as context, note that (with NMOS) if we (i) apply a suitable voltage to the gate & (ii) then apply a suitable voltage between source and drain, current will flow.

Moore's Law

- "Cramming more components onto integrated circuits."

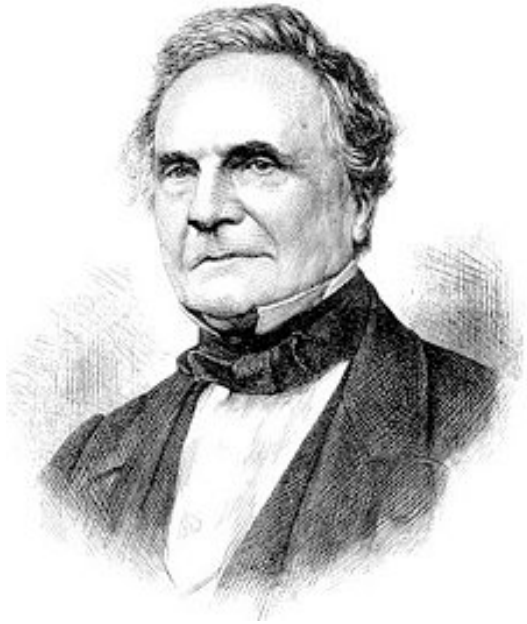
  - G.E. Moore, Electronics 1965

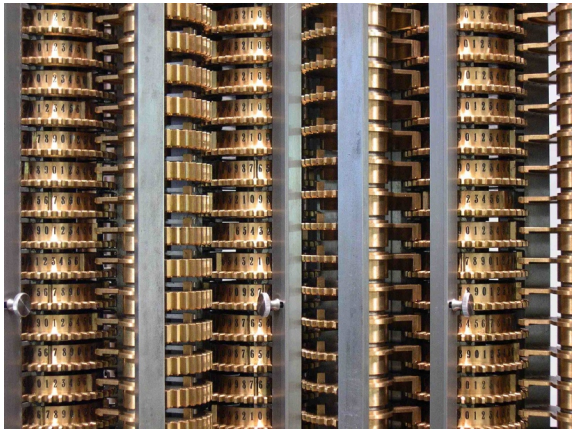
- Observation: DRAM transistor density doubles annually
  - Became known as "Moore's Law"
  - Actually, a bit off:
    - Density doubles every 18 months
    - (in 1965 they only had 4 data points!)

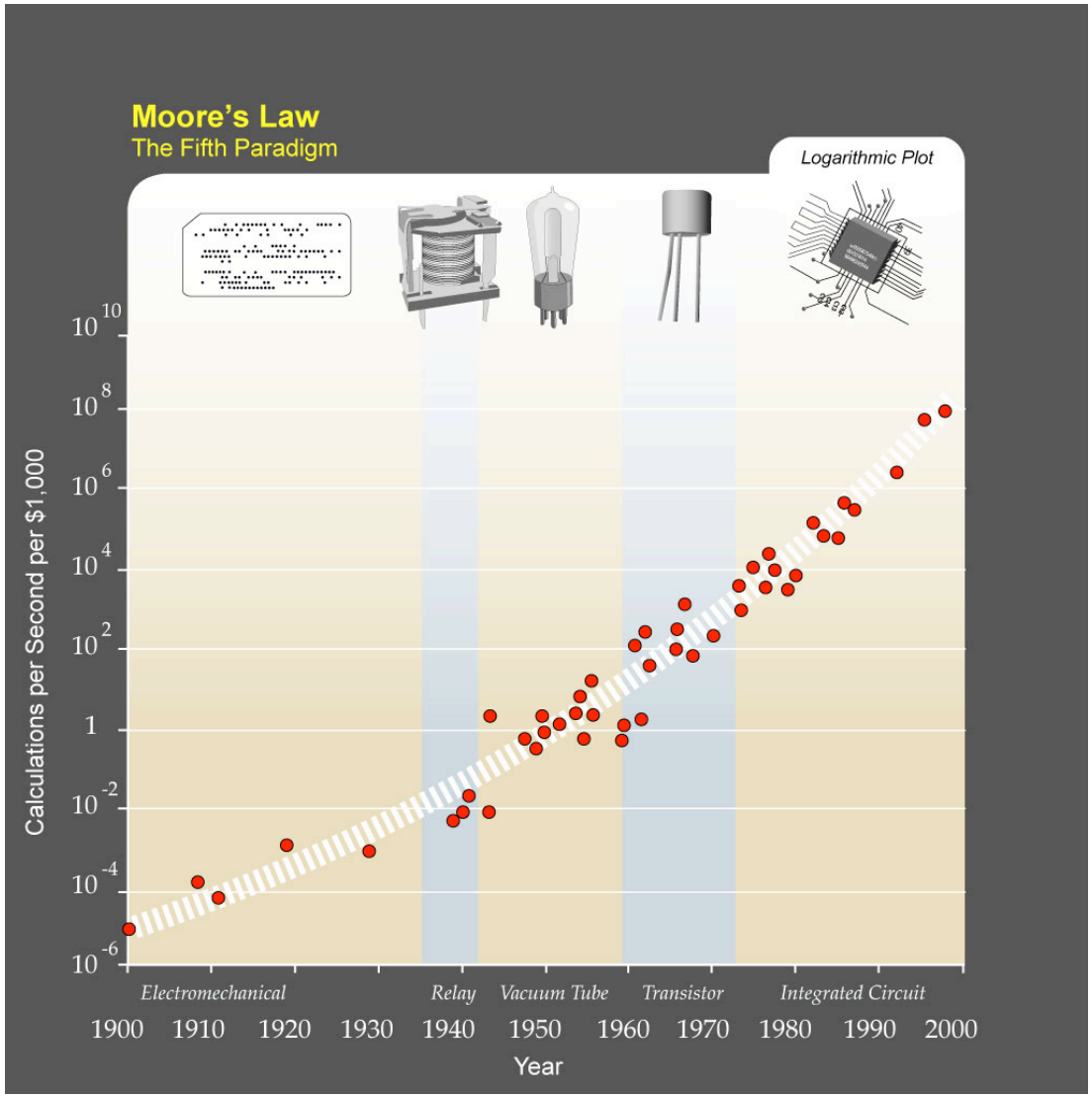
- Corollaries:
  - Cost per transistor halves annually (18 months)
  - Power per transistor decreases with scaling
  - Speed increases with scaling
  - Reliability increases with scaling
    - Of course, it depends on how small you try to make things
      - » (I.e. no exponential lasts forever)

Remember these!

Moore's Law






Moore's Law
The Fifth Paradigm
Logarithmic Plot
Calculations per Second per $1,000
10^{10}
10^{8}
10^{6}
10^{4}
10^{2}
1
10^{-2}
10^{-4}
10^{-6}
Electromechanical
Relay
Vacuum Tube
Transistor
Integrated Circuit
1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000
Year

Moore's Law

- Moore's Curve is a self-fulfilling prophecy
  - 2X every 2 years means ~3% per month
    - I.e. ((1 X 1.03) * 1.03)*1.03… 24 times = ~2
  - Can use 3% per month to judge performance features
  - If feature adds 9 months to schedule…it should add at least 30% to performance
    - ($1.03^9 = 1.30$ ⇨ 30%)

A bit on device performance...

- One way to think about switching time:
  - Charge is carried by electrons
  - Carrier velocity is proportional to the lateral E-field between source and drain
    - i.e. $v = mE$
      - m = carrier mobility (and can be though of as a constant)
  - Electric field defined as: $E = V_{ds}/L$
  - Time for charge to cross channel = length/speed
    - (i.e. meters / (meters/s) = seconds)
    - $= L/v$
    - $= L/(mE)$
    - $= L/(m*(V_{ds}/L))$
    - $= L^2/(mV_{ds})$

Thus, to make a device faster, we want to either increase V_{ds} or decrease feature sizes (i.e. L)

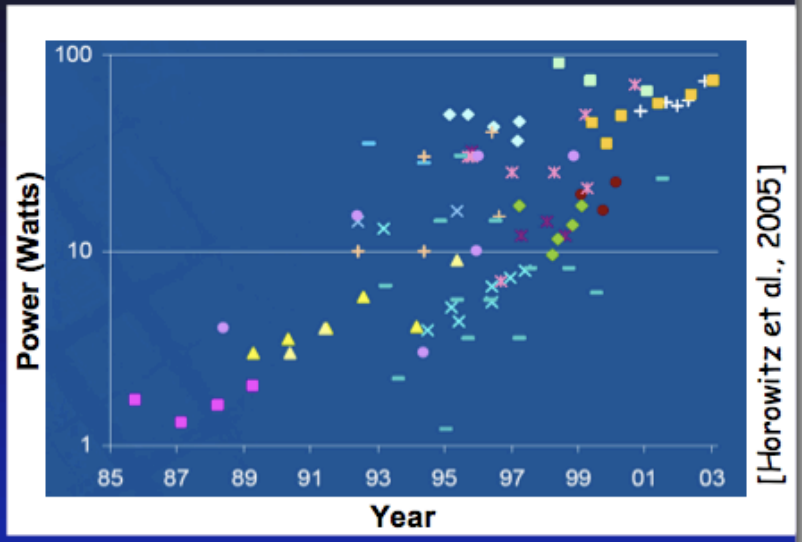
Some more important relationships

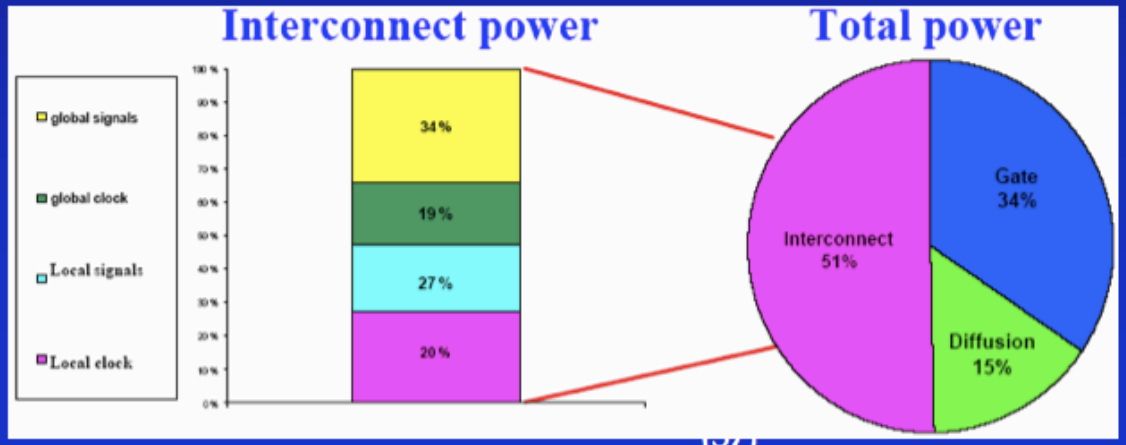
- **What about power?**
  - **First, need to quickly discuss equation for capacitance:**
    - **$C_L = (e_{ox}WL)/d$**
      - e_{ox} = dielectric, WL = parallel plate area, d = distance between gate and substrate
  - **Then, dynamic power becomes:**
    - **$P_{dyn} = C_L V_{dd}^2 f_{0\text{-}1}$**
      - Dynamic power is a function of the frequency of 0 to 1 or 1 to 0 transitions (as this involves the movement of charge)
        - » Note frequency in this context is NOT clock frequency
      - Note that as W and L scale, C_L decreases which in turn will cause a decrease in P_{dyn}.
      - Note that while an increase in V_{dd} will *decrease* switching time, it will also cause a quadratic *increase* in dynamic power.

Interconnect plays a role too...

Interconnect Power Dissipation

- Power dissipation is arguably the most critical problem in high-performance chip design
- Over last two decades microprocessor power dissipation grows exponentially and primary contribution from interconnects





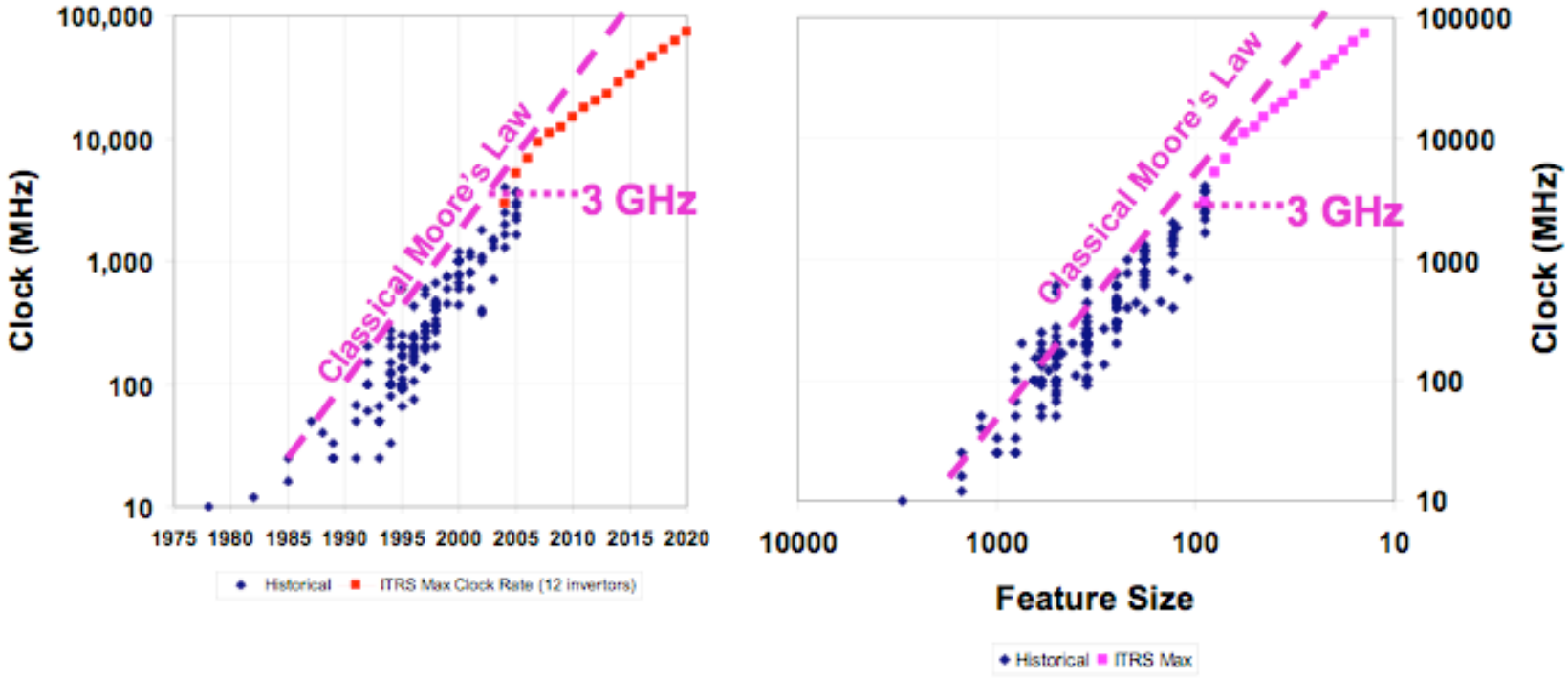
Interconnect responsible for 50% of dynamic power dissipation [Magen et al., 2004]

COLUMBIA UNIVERSITY

A funny thing happened on the way to 45 nm

- Speed increases with scaling...

Remember these!

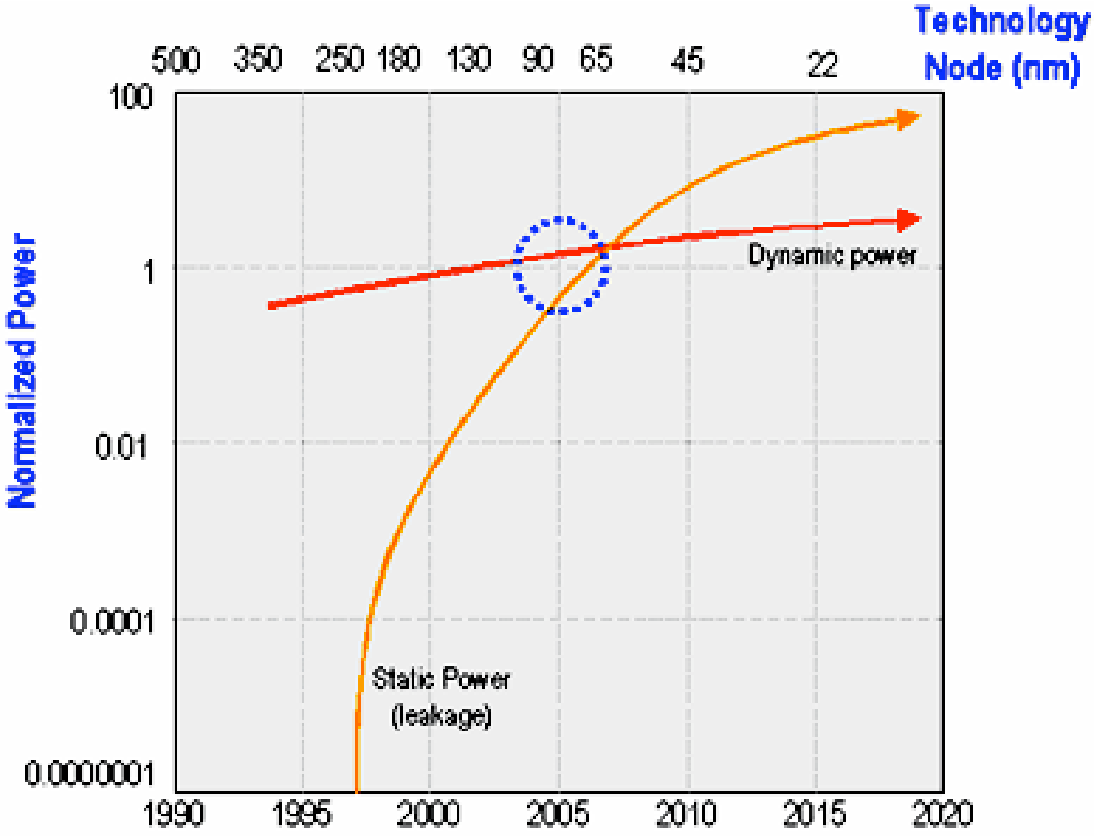


2005 projection was for 5.2 GHz - and we didn't make it in production. Further, we're still stuck at 3+ GHz in production.

A funny thing happened on the way to 45 nm

- Power decreases with scaling...

Remember these!



A funny thing happened on the way to 45 nm

- Power decreases with scaling...

Remember these!

This is not the trend that you want to see:

Decreasing Supply Voltage, Increasing Power

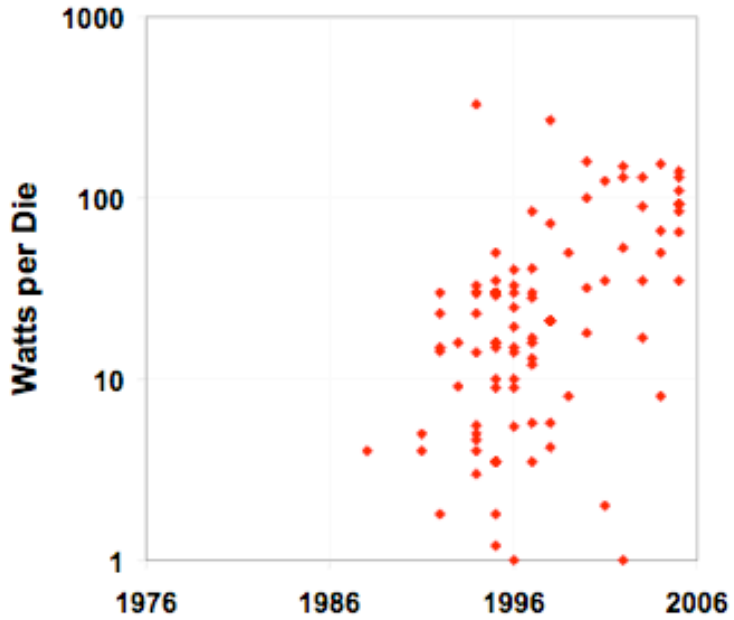
YEAR	2004	2007	2010	2013	2016
TECHNOLOGY	90 nm	65 nm	45 nm	32 nm	22 nm
CHIP SIZE	550 mm^2	550 mm^2	550 mm^2	550 mm^2	550 mm^2
NUMBER OF TRANSISTORS (LOGIC)	553 M	1 Billion	2 Billion	4.5 Billion	8.5 Billion
DRAM CAPACITY	1.0 Gbits	2.0 Gbits	4.3 Gbits	8.5 Gbits	35 Gbits
MAXIMUM CLOCK FREQUENCY	4.1 GHz	9.3 GHz	15 GHz	23 GHz	40 GHz
MINIMUM SUPPLY VOLTAGE	0.9 V	0.8 V	0.7 V	0.6 V	0.5 V
MAXIMUM POWER DISSIPATION	150 W	190 W	200 W	200 W	200 W
MAXIMUM NUMBER OF I/O PINS	3000	4000	4000	5300	7000

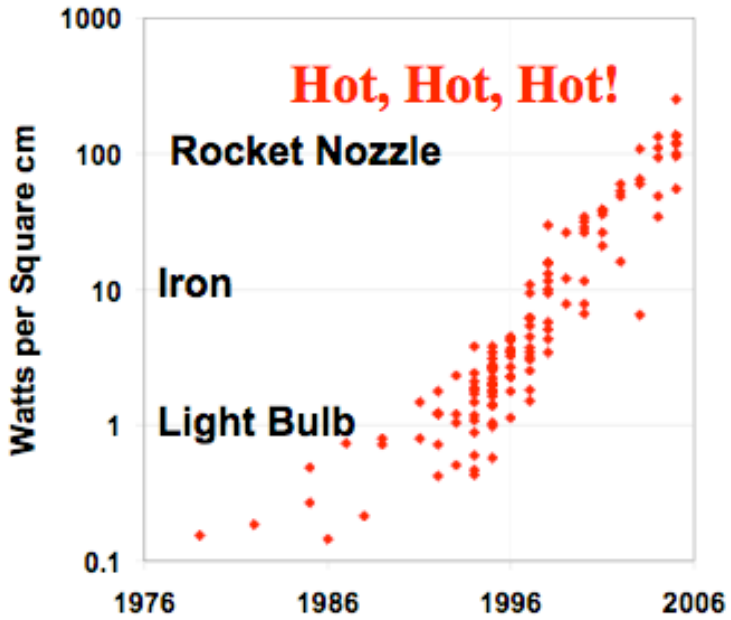
A funny thing happened on the way to 45 nm

- Speed increases with scaling...
- Power decreases with scaling...

Remember these!

Why the clock flattening? POWER!!!!

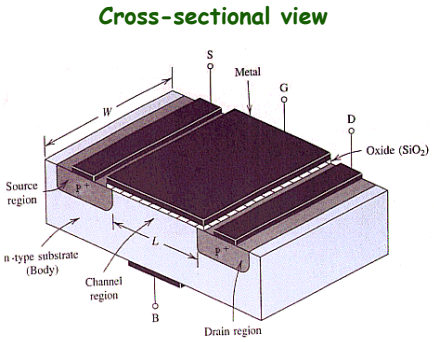


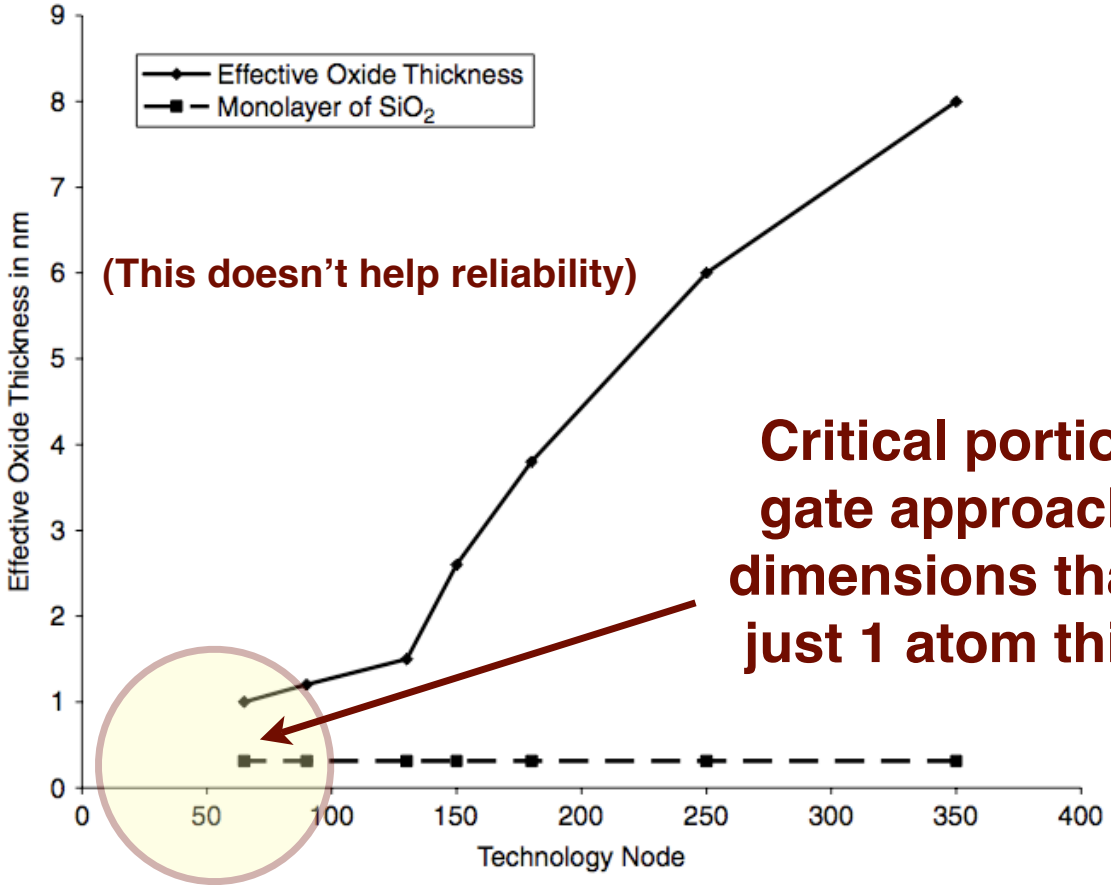


A funny thing happened on the way to 45 nm

- Reliability increases with scaling... Remember these!

One quick example:



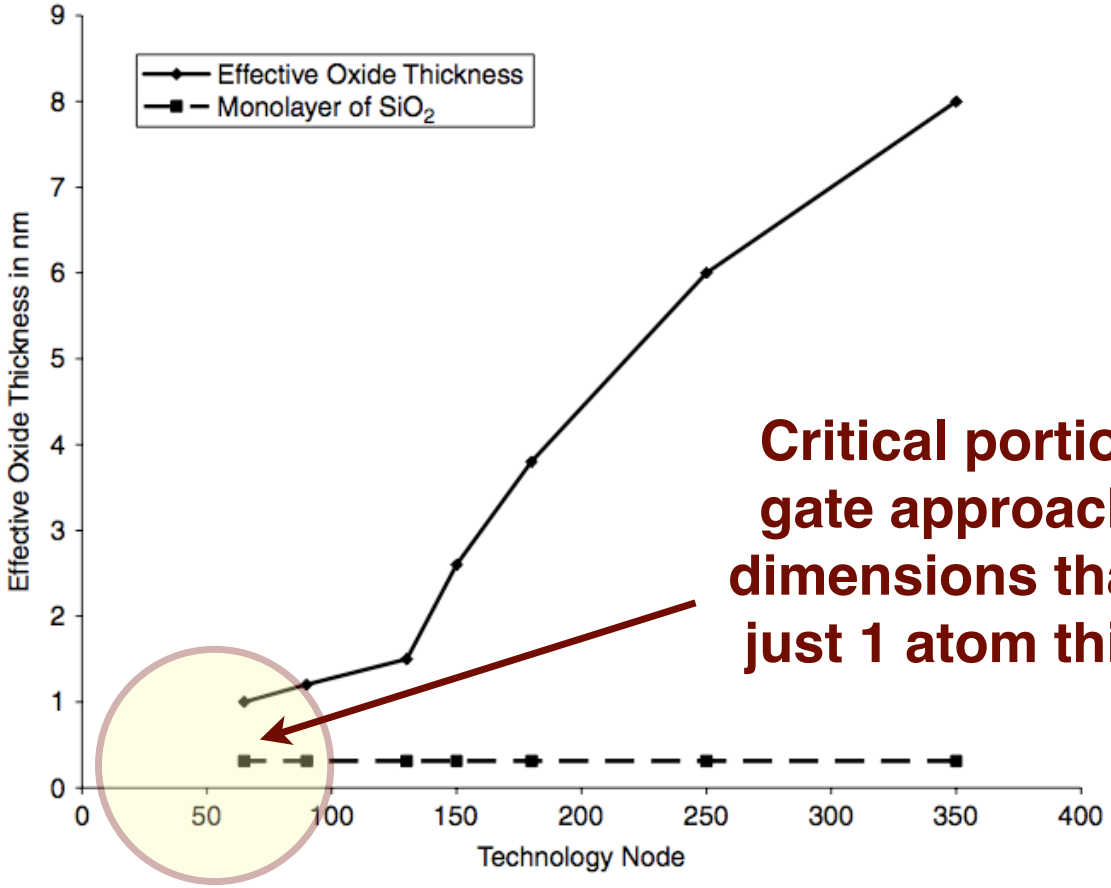


A funny thing happened on the way to 45 nm

- Power decreases with scaling...

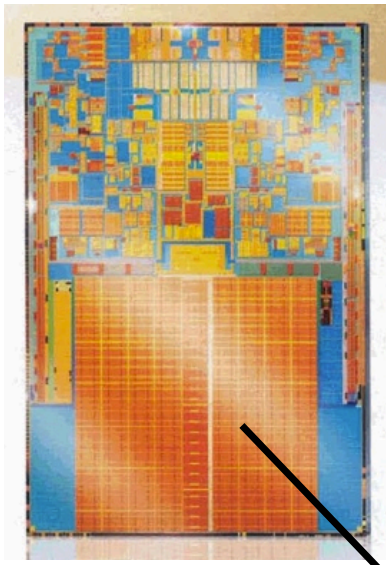
Remember these!

At such small dimensions, quantum mechanical effects take over and devices no longer turn off well.



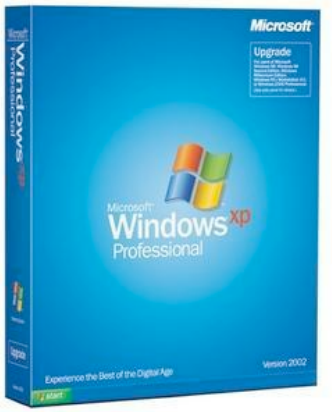
Critical portion of gate approaching dimensions that are just 1 atom thick!!!

Transistors used for memory too...



Saw earlier, that transistors used for *on chip* memory too...

Problem: Program, data = bigger + power problem isn't going away...



- Why? Put faster memory closer to processing logic...
  - SRAM (logic): density +25%, speed +20%
  - DRAM (memory): density +60%, speed +4%
  - Disk (magnetic): density +25%, speed +4%
  - Network interface: 10 Mb/s ⇨ 100 Mb/s ⇨ 1 Gb/s

High art meets high-tech.

Lincoln's latest project, titled "CUBE," is a 10' x 10' translucent structure outfitted with video cameras, uniquely combining sculpture, portraiture and architecture. With **Intel® Centrino® processor technology** inside, a notebook becomes many other things as well — portable studio, canvas, inspiration tool.

Top 5 Must-Haves

☐ POWERFUL PROCESSOR
A portrait of performance. "My generative portraits are demanding on the processors in my laptop, as they continuously manipulate video," says Lincoln. Thankfully, the **dual-core performance** of Intel Centrino processor technology can handle intensive tasks with flying colors.

☐ DIZZYING TRANSFER SPEEDS
Art (at 30 frames per second). Data transferring up to 20% faster[1] allows Lincoln to store footage from 24 video cameras with lightning speed

☐ HIGH-SPEED WIRELESS
Always Connected. With up to **twice the range** and **5x the speed** when connected to a Wireless N home network,[2] Lincoln can download music or shop for art books anywhere, anytime.

☐ ENHANCED VIDEO
High-def (redefined). Lincoln can view his generative portraits with "gallery-like" clarity, thanks to **stunning multimedia performance**, for a super-enhanced high-def video experience.

☐ [illegible]
The power of art. Lincoln's infinitely reconfiguring images are ultimately presented on a plasma screen powered by his computer — so wasting power is not an option. Thanks to **Intel's exclusive power-saving features**, he conserves energy by using it only when he needs it.

Deeper. Richer. Faster.
Log on to drivenbywhatsinside.com for access to exclusive multimedia content to keep you up-to-date on the latest tec trends — faster. To take advantage of this high-tech, multimedia material, make sure your computer has **Intel Centrino processor technology**.

intel Centrino

Solution?

Motivation:
Processor complexity is good enough, transistor sizes scale, we can slow processors down, manage power, and get performance from...

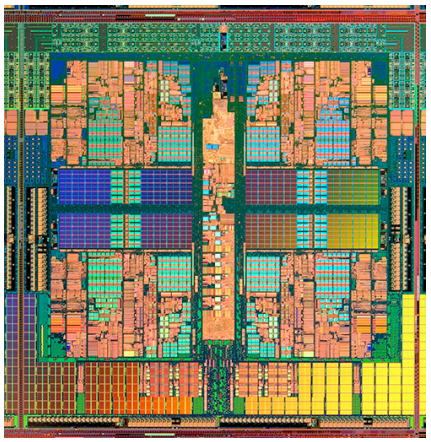
Parallelism

Top 5 Must-Haves

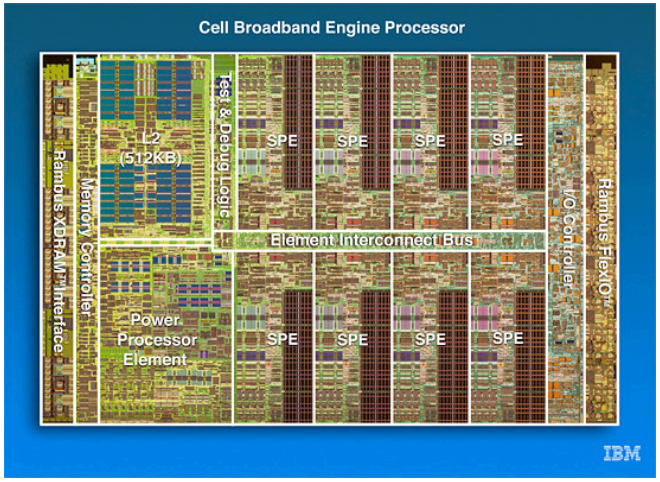
☐ POWERFUL PROCESSOR
A portrait of performance. "My generative portraits are demanding on the processors in my laptop, as they continuously manipulate video," says Lincoln. Thankfully, the **dual-core performance** of Intel Centrino processor technology can handle intensive tasks with flying colors.

This idea has been extended...

Quad core chips...



7, 8, and 9 core chips...



When will it stop?

Practical problems must be addressed!

So, what are the goals of this course?

- At the end of the semester, you should be able to...
  - ...explain and articulate why modern microprocessors now have more than 1 core...
- Why?
  - For 8, 16 core chips to be practical, we have to be able to *use them*
    - Students in this class should go on to play a role in making such chips useful...

So, what are the goals of this course?

- At the end of the semester, you should be able to...
  - Apply fundamental knowledge about single core machines, dual core machines, performance metrics, etc. to design a microprocessor such that it (a) meets a target set of performance goals and (b) is realistically implementable.

A few notes about Lecture 02 (guest lecture) (print out notes)

Now, let's look at the syllabus