Nvidia CEO Jen-Hsun Huang said: “whatever capability you think you have today, it’s nothing compared to what you’re going to have in a couple of years … due to supercomputers in the cloud”. Now, cloud computing does what you could do on your PC. Imagine 40,000 times that!

www.theregister.co.uk/2010/09/24/huang_muses_at_gtc/
Happy Confucius Day!

Knowledge is recognizing what you know and what you don't.

I hear and I forget.
I see and I remember.
I do and I understand.
Concurrency & Parallelism, 10 mi up...

**Intra-computer**
- Today’s lecture
- Multiple computing “helpers” are cores within one machine
- Aka “multi-core”
  - Although GPU parallelism is also “intra-computer”

**Inter-computer**
- Week 12’s lectures
- Multiple computing “helpers” are different machines
- Aka “distributed computing”
  - Grid & cluster computing
Anatomy: 5 components of any Computer
Anatomy: 5 components of any Computer

John von Neumann invented this architecture

Computer

Processor
- Control ("brain")
- Datapath ("brawn")

Memory

Devices
- Input
- Output

What causes the most headaches for SW and HW designers with multi-core computing?

a) Control
b) Datapath
c) Memory
d) Input
e) Output
But what is INSIDE a Processor?

Processor

- Control
  - (“brain”)
- Datapath
  - (“brawn”)

Garcia, Fall 2010
But what is INSIDE a Processor?

- Primarily Crystalline Silicon
- 1 mm – 25 mm on a side
- 2009 “feature size” (aka process) 
  \[ \sim 45 \text{ nm} = 45 \times 10^{-9} \text{ m} \] 
  (then 32, 22, and 16 [by yr 2013])
- 100 - 1000M transistors
- 3 - 10 conductive layers
- “CMOS” (complementary metal oxide semiconductor) - most common

- Package provides:
  - spreading of chip-level signal paths to board-level
  - heat dissipation.
- Ceramic or plastic with gold wires.
Moore's Law

Predicts: 2X Transistors / chip every 2 years

What is this “curve”?  
- a) Constant  
- b) Linear  
- c) Quadratic  
- d) Cubic  
- e) Exponential

Moore's law predicts that the number of transistors on an integrated circuit (IC) doubles every two years. The curve shows the historical trend of transistor count on chips from 1971 to 2008. The options a) to e) represent different types of curves that could fit this data. The correct answer is:  
- e) Exponential
Moore's Law and related curves

Data partially collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond

UC Berkeley CS10 “The Beauty and Joy of Computing”: Concurrency (9)
Power Density Prediction circa 2000

Source: S. Borkar (Intel)
### Going Multi-core Helps Energy Efficiency

- **Power of typical integrated circuit ~ C V^2 f**
  - C = Capacitance, how well it “stores” a charge
  - V = Voltage
  - f = frequency. I.e., how fast clock is (e.g., 3 GHz)

<table>
<thead>
<tr>
<th></th>
<th>Core</th>
<th>Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voltage</td>
<td>1</td>
<td>-15%</td>
</tr>
<tr>
<td>Freq</td>
<td>1</td>
<td>-15%</td>
</tr>
<tr>
<td>Area</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Power</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Perf</td>
<td>1</td>
<td>~1.8</td>
</tr>
</tbody>
</table>

Activity Monitor (on the lab Macs) shows how active your cores are

---

William Holt, HOT Chips 2005

UC Berkeley CS10 “The Beauty and Joy of Computing” : Concurrency (12)
Energy & Power Considerations

\[
\text{Power} = \frac{\text{Energy}}{\text{Second}} = \frac{\text{Energy}}{\text{Op}} \times \frac{\text{Ops}}{\text{Second}}
\]

- **Power**
  - Chip Packaging
  - Chip Cooling
  - System Noise
  - Case Temperature
  - Data-Center Air Conditioning

- **Energy**
  - Battery Life
  - Electricity Bill
  - Mobile Device Weight

**Courtesies:**
- Chris Batten
- UC Berkeley CS10 “The Beauty and Joy of Computing” : *Concurrency* (13)
Parallelism again? What’s different this time?

“This shift toward increasing parallelism is not a triumphant stride forward based on breakthroughs in novel software and architectures for parallelism; instead, this plunge into parallelism is actually a retreat from even greater challenges that thwart efficient silicon implementation of traditional uniprocessor architectures.”

– Berkeley View, December 2006

- HW/SW Industry bet its future that breakthroughs will appear before it’s too late
Background: Threads

- **Thread** stands for “thread of execution”, is a single stream of instructions
  - A program / process can split, or *fork* itself into separate threads, which can (in theory) execute simultaneously.
  - An easy way to describe/think about parallelism
- A single CPU can execute many threads by **Time Division Multiplexing**
- **Multithreading** is running multiple threads through the same hardware
Applications can almost never be completely parallelized; some serial code remains.

\[ \text{Speedup}(P) = \frac{\text{Time}(1)}{\text{Time}(P)} \leq \frac{1}{s + (1-s)/P}, \text{ and as } P \to \infty \leq \frac{1}{s} \]

Even if the parallel portion of your application speeds up perfectly, your performance may be limited by the sequential portion.
Speedup Issues: Overhead

- Even assuming no sequential portion, there’s…
  - Time to think how to divide the problem up
  - Time to hand out small “work units” to workers
  - All workers may not work equally fast
  - Some workers may fail
  - There may be contention for shared resources
  - Workers could overwriting each others’ answers
  - You may have to wait until the last worker returns to proceed (the slowest / weakest link problem)
  - There’s time to put the data back together in a way that looks as if it were done by one
Life in a multi-core world…

- This “sea change” to multi-core parallelism means that the computing community has to rethink:
  a) Languages
  b) Architectures
  c) Algorithms
  d) Data Structures
  e) All of the above
But parallel programming is hard!

- What if two people were calling withdraw at the same time?
  - E.g., balance=100 and two withdraw 75 each
  - Can anyone see what the problem could be?
  - This is a race condition

- In most languages, this is a problem.
  - In Scratch, the system doesn’t let two of these run at once.
Another concurrency problem ... deadlock!

- Two people need to draw a graph but there is only one pencil and one ruler.
  - One grabs the pencil
  - One grabs the ruler
  - Neither release what they hold, waiting for the other to release

- Livelock also possible
  - Movement, no progress
  - Dan and Luke demo

en.wikipedia.org/wiki/Deadlock
Summary

 “Sea change” of computing because of inability to cool CPUs means we’re now in multi-core world
 This brave new world offers lots of potential for innovation by computing professionals, but challenges persist