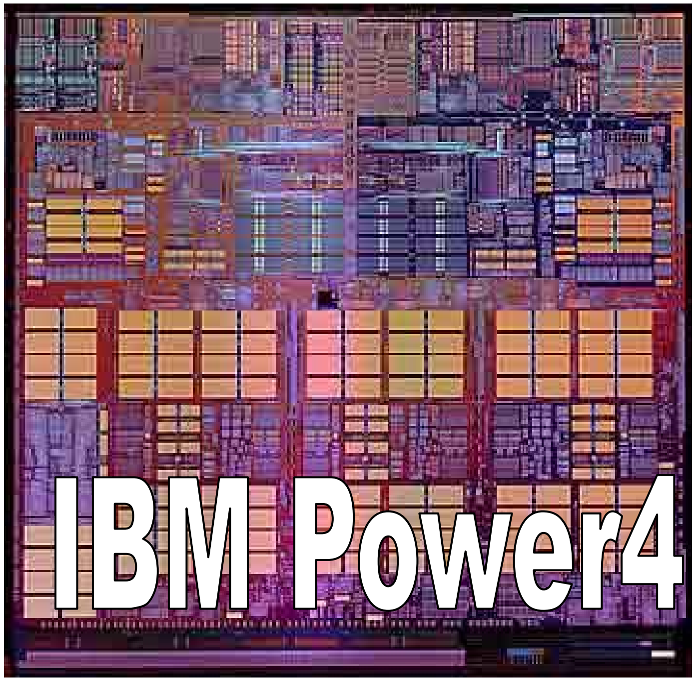
18-200

Introduction to Computer Architecture at Carnegie Mellon **Or** What are we doing with all these transistors?



Fall 2004

Prof. Babak Falsafi

© 2004 Babak Falsafi

Babak Falsafi

Associate Professor
Department of ECE & CS
CALCM
CSSI
http://www.ece.cmu.edu/~babak

Courses:

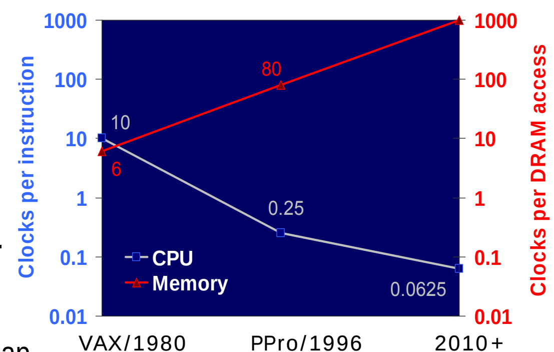
18-240 Fundamentals of Computer Engineering
18-447: Introduction to Computer Architecture
18-741: Advanced Computer Architecture
18-747: Advanced Topics in Microarchitecture

High-Perf. Memory Systems

Tempest *(w/ Ailamaki)*

- Streaming memory engine
- Correlate temporal data & insts.
- Predict "when" to stream
- Bridge the CPU/memory perf. gap

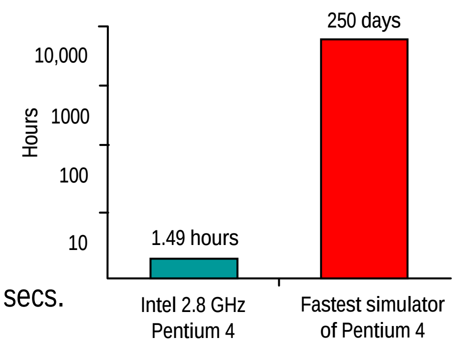
Processor/memory performance gap will continue to increase!



SimFlex *(w/ Hoe & Nowatzyk)*

- Fast & Accurate Simulation
- Using Statistical sampling
- Reduce time/benchmark from days to secs.

Benchmark (SPEC 2K) Execution Time



Full-workload simulation is not tractable!

Nanoscale CMOS Systems

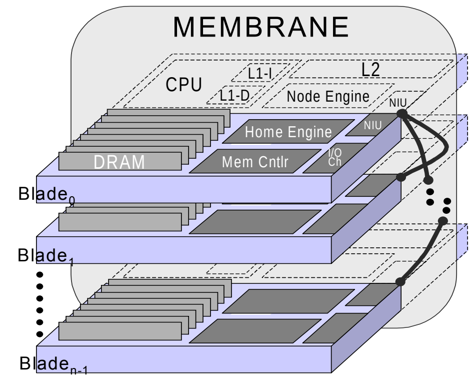
TRUSS *(w/ Hoe & Nowatzyk)*

Failure trends:

- High chip density -> more error-prone
- High hard & transient failure rate
- Downtime cost for brokerage operations: $6,450,000/h [Patterson, ROC keynote]

TRUSS = "RAID" for Computing

- Non-Stop Scalable Servers



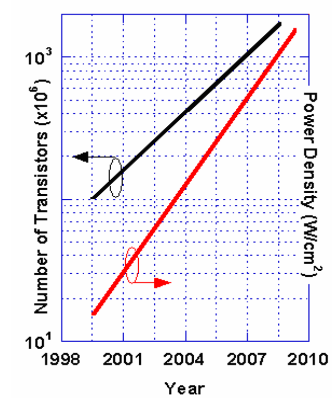
Chiller *(w/ Asheghi)*

Dynamic HotSpot management

- Average heat much lower than peak
- But, designs target worst-case heat -> expensive + unreliable

Chiller processors:

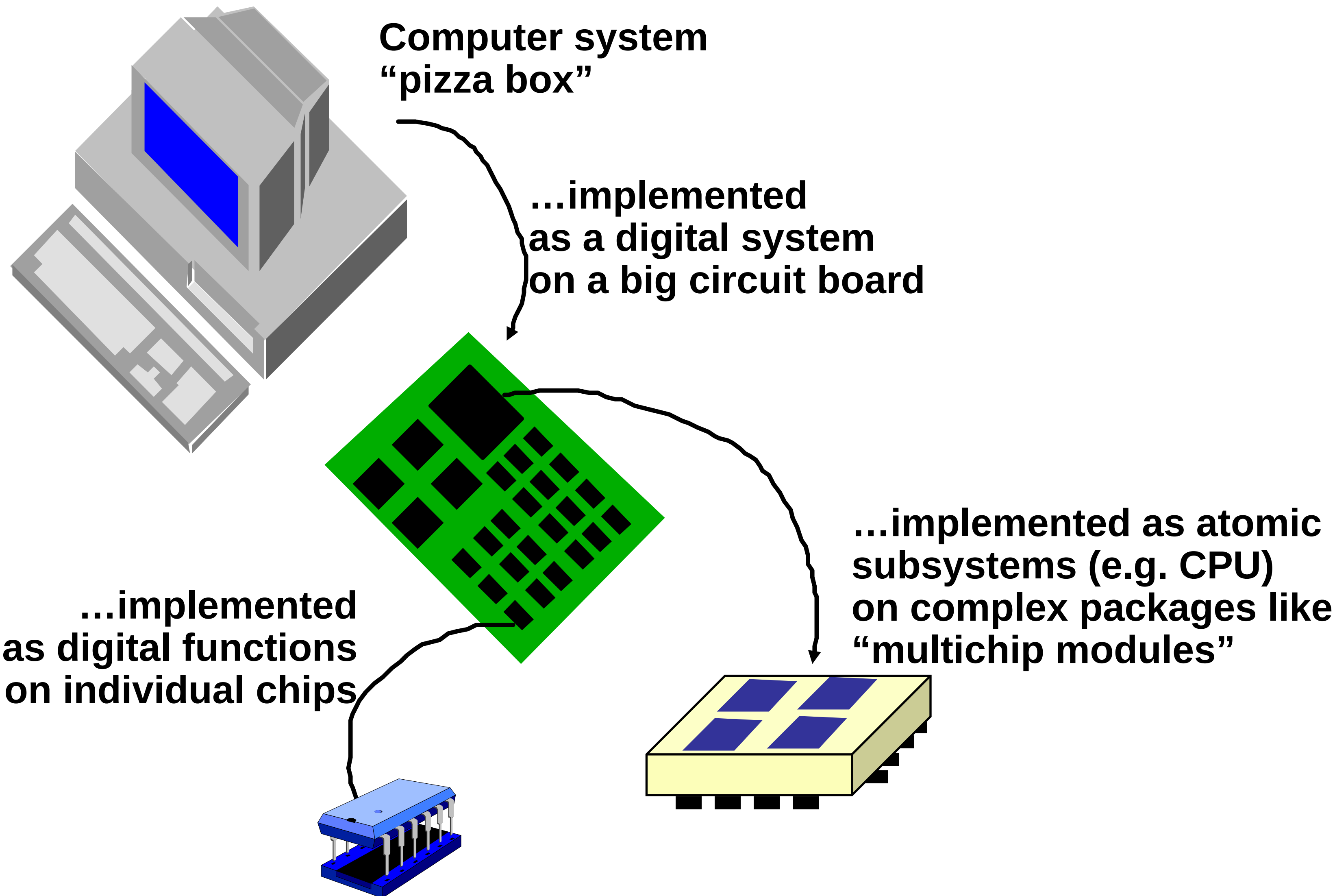
- Identify HotSpots using nanosensors
- Chill on the fly



© 2004 Babak Falsafi

babak@cmu.edu

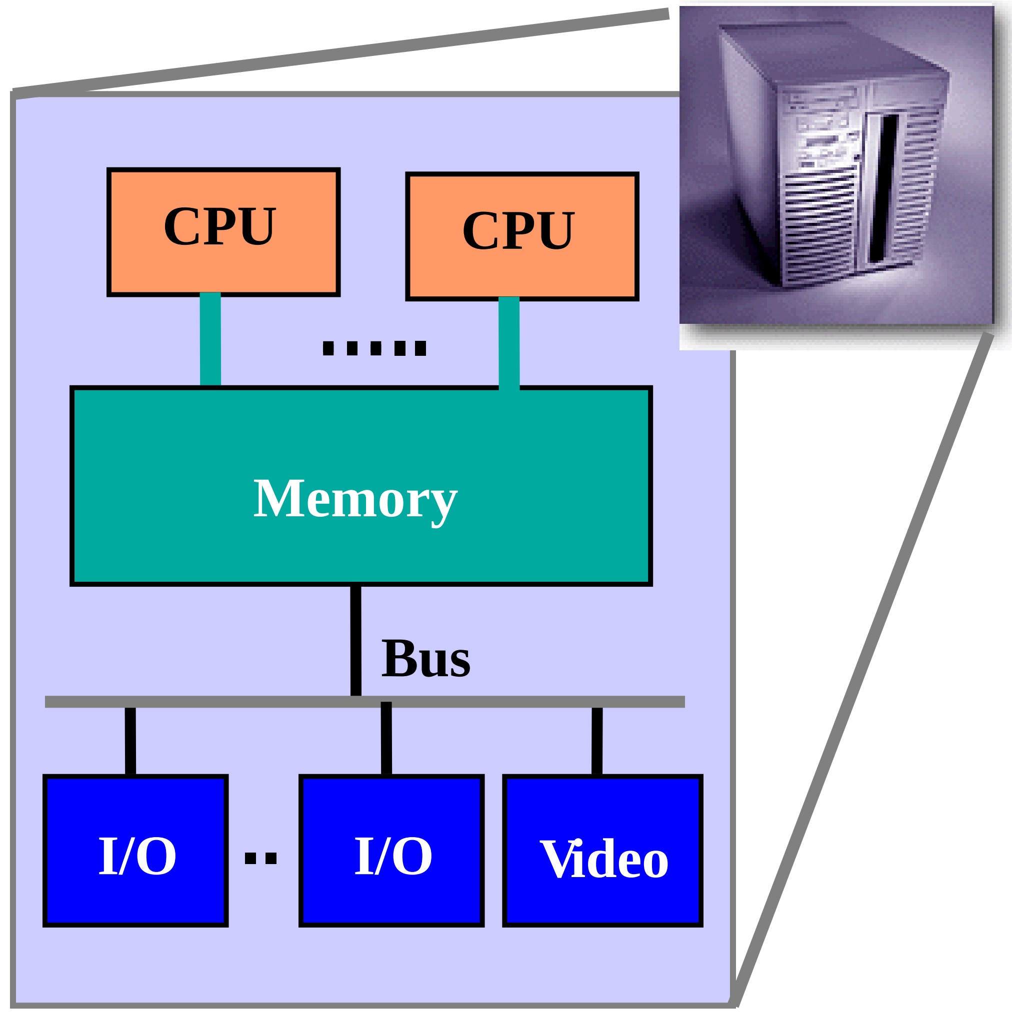
Computers are Digital Systems



© 2004 Babak Falsafi

What's Inside a Computer

- Processor(s), Microprocessor(s), or CPU(s)
- Memory subsystem
- I/O subsystem
  - network, disk drives, keyboard, mouse, etc.



© 2004 Babak Falsafi

What is Computer Architecture?

- The science and art of selecting and interconnecting hardware components to create a computer that meets functional, performance and cost goals.

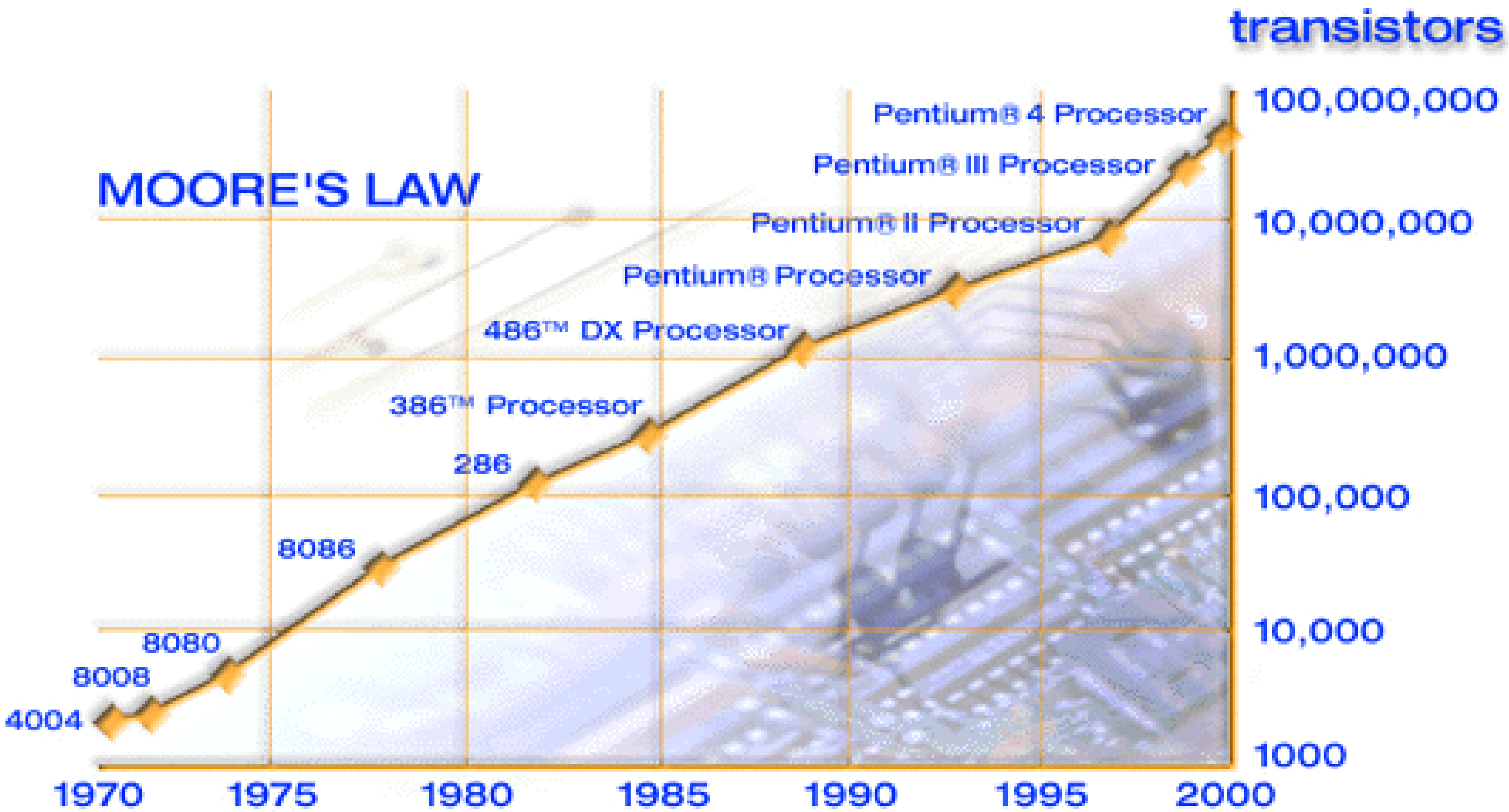
Selection process evolves because

- Technology changes

→ Enables new applications

© 2004 Babak Falsafi

Why Study Computer Architecture?



http://www.intel.com/research/silicon/mooreslaw.htm

Besides More Transistors….

Other aspects also improve

Technology	Annual Improvement
Transistor speed	20%-25%
Memory density	60%
Memory speed	4%
Disk density	25%
Disk speed	4%

© 2004 Babak Falsafi

Processor Performance

"Unmatched by any other industry" [John Crawford, Intel Fellow, 1993]

Doubling every 18 months (1982-1996): total of 800X

- Cars travel at 44,000 MPH; get 16,000 miles/gal.
- Air travel: L.A. to N.Y. in 22 seconds (MACH 800)
- Wheat yield: 80,000 bushels per acre

Doubling every 24 months (1970-1996): total of 9,000X

- Cars travel at 600,000 MPH; get 150,000 miles/gal.
- Air travel: L.A. to N.Y. in 2 seconds (MACH 9,000)
- Wheat yield: 900,000 bushels per acre

Exponential effect

© 2004 Babak Falsafi

User Requirements Change Too

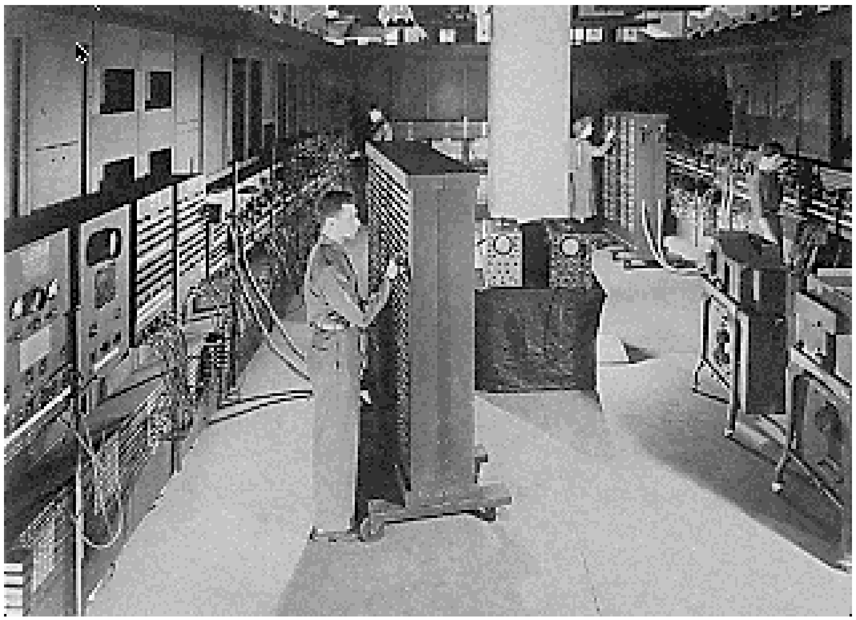
- Types of applications today
  - Scientific
    - Weather prediction, crash analysis, earthquake analysis, medical imaging, imaging of the earth (searching for oil)
  - Business
    - database, data mining, video
  - General purpose
    - Microsoft Word, Excel
  - Real-time
    - automated control systems
  - Games
    - Nintendo
  - Mobile

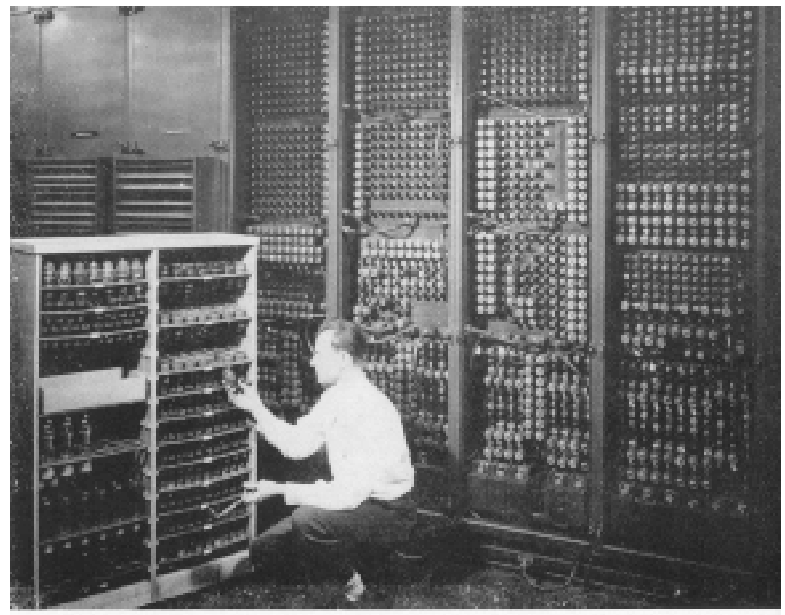
- Tomorrow: you can think of one before leaving class today.....

© 2004 Babak Falsafi

A Historical Perspective

- In the beginning…Eniac





Replacing a bad tube meant checking among ENIAC's 19,000 possibilities.

5,000 additions in one second

© 2004 Babak Falsafi

Eniac

- Built at the University of Pennsylvania
  - Lt Gillon, Eckert and Mauchley
- Initial contract for $61,700, June 1943
  - eventually cost $486,804.22, in 1946
  - Accumulator deployed in Jun 1944
  - Accumulator, multiplier, divide and square root and 3 portable function tables completed in Fall 1945
- 200 μsecond cycle time for 1 add
- Internals
  - 19K vacuum tubes, 1.5K relays, 100K's of resistors, and inductors;
30 separate units; forced air cooling; Multiply's in base 10; just like a human
  - Originally, no internal memory --> programmed w/cables and switches
- Designed to compute firing tables
  - Differential equations of motion to compute trajectory in 15 seconds (same amount of computation took a human 20 hours)
- Power = 200K Watts

© 2004 Babak Falsafi

Four Decades of Microprocessor

The Decade of the 1970's "Microprocessors"

- Single-Chip Microprocessors
- Personal Computers (PC)

The Decade of the 1980's "Quantitative Architecture"

- Instruction Pipelining
- Fast Cache Memories
- Workstations

The Decade of the 1990's "Instruction-Level Parallelism"

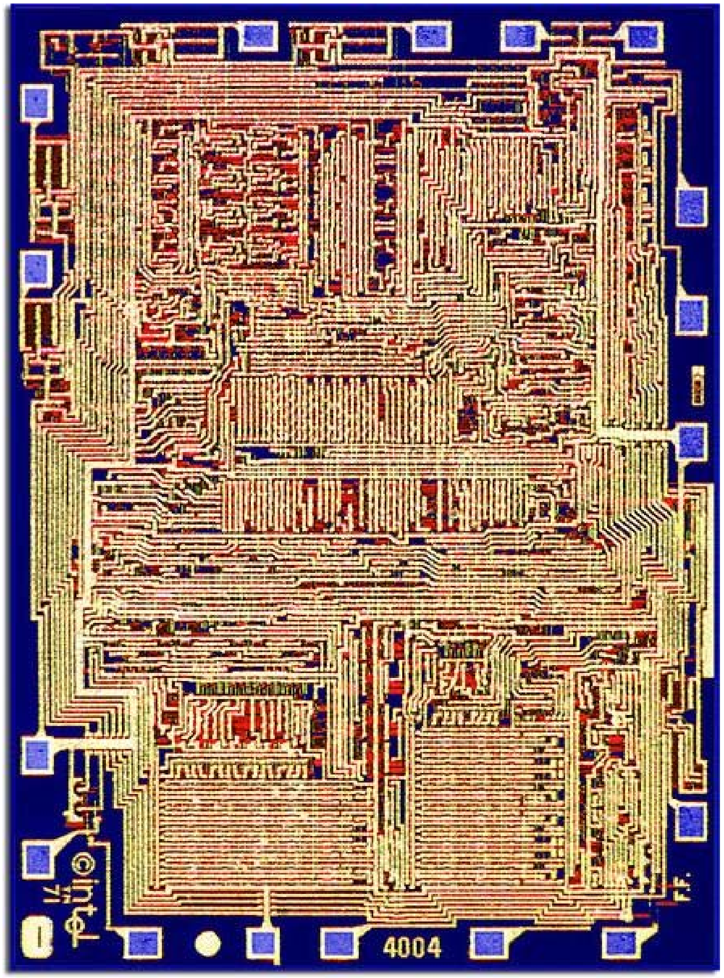
- Superscalar Processors
- Low-Cost Desktop Supercomputing

The Decade of the 2000's

You will learn in 18-347/18-741/18-742/18-747

© 2004 Babak Falsafi

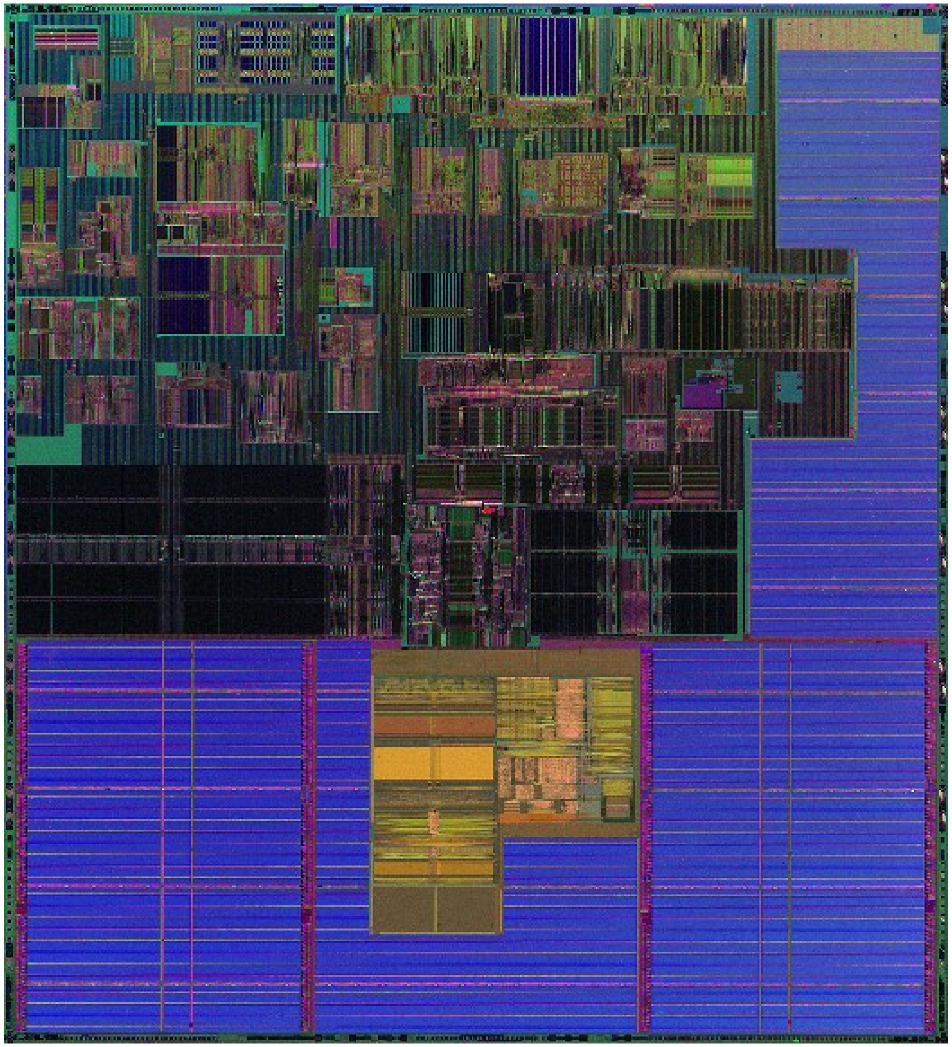
Intel 4004, circa 1971



- **The first single chip CPU**
  - 4-bit processor for a calculator.
  - 1K data memory
  - 4K program memory
  - 2,300 transistors
  - 16-pin DIP package
  - 740kHz (eight clock cycles per CPU cycle of 10.8 microseconds)
  - ~100K OPs per second

Molecular Expressions: Chipshots

Intel Itanium 2, circa 2002



http://cpus.hp.com/images/die_photos/McKinley_die.jpg

- **Performance leader in floating-point apps**
  - 64-bit processor
  - 3 MByte on-chip memory!!
  - 221 million transistor
  - 1 GHz, issue up to 8 instruction per cycle

In ~30 years, about 100,000 fold growth in transistor count and performance!

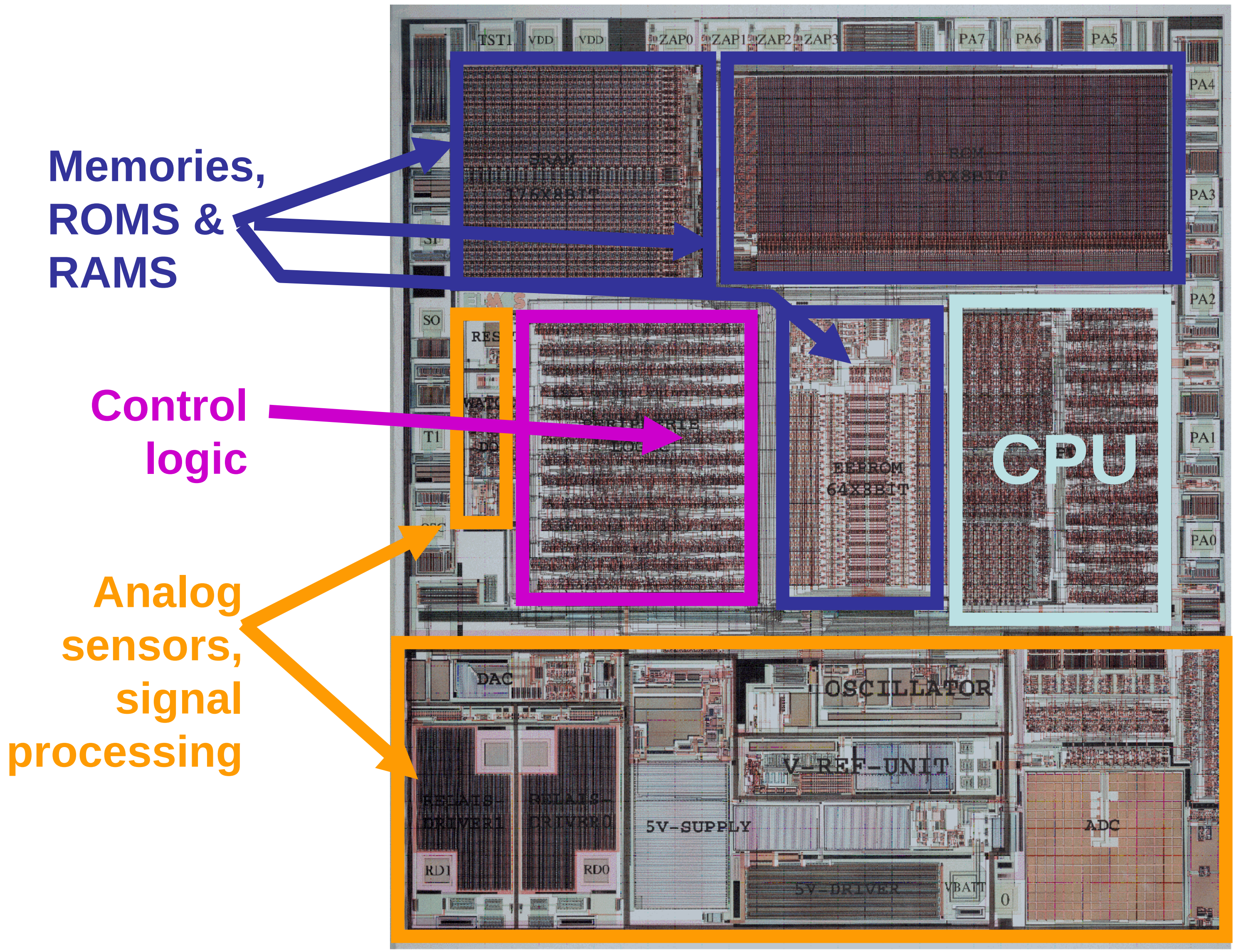
Where Are Microprocessors?

- Right now...
  - Supercomputers, workstations, PCs
  - Set-Top boxes
  - DVD players, CD players
  - Cars (many per car)
  - Modems, network cards
  - Toasters, microwave ovens, fridges

- Soon...
  - Your clothing, your glasses, your jewelry...
  - *Everywhere*

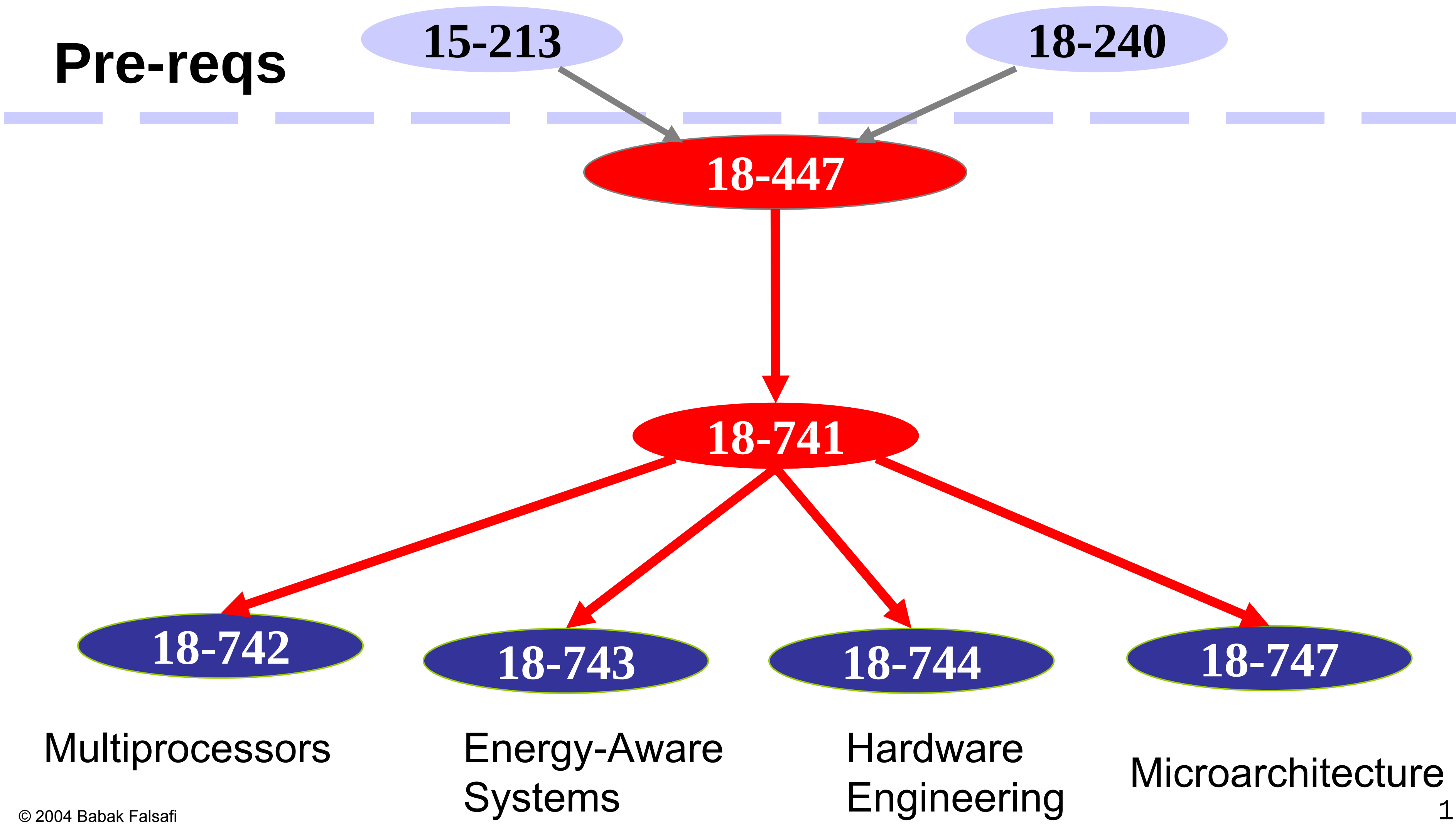
© 2004 Babak Falsafi

"Non-computer" ICs Have CPUs Too



© 2004 Babak Falsafi

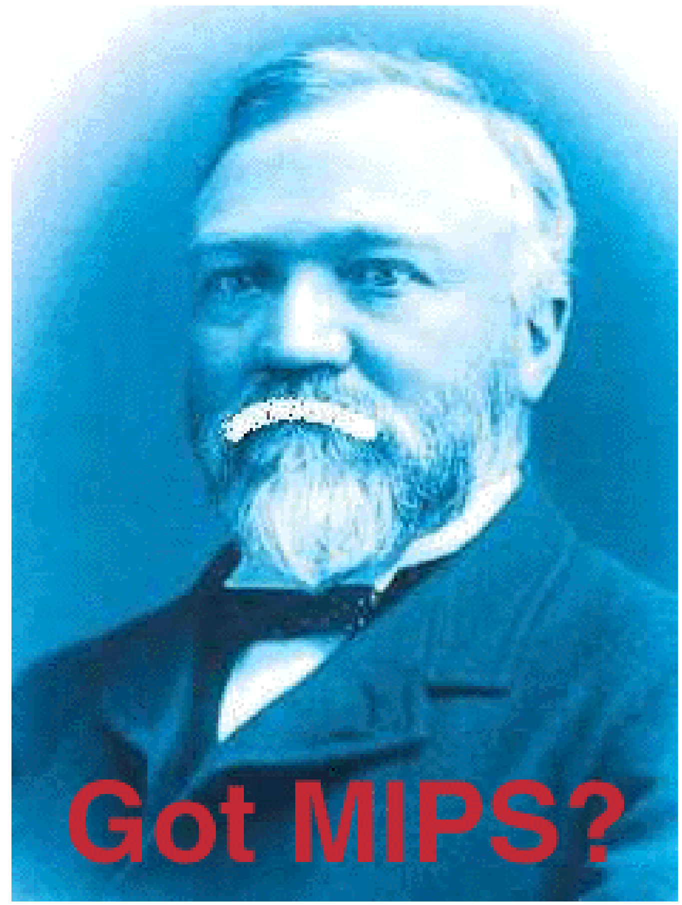
Computer Architecture @ CMU: Curriculum



© 2004 Babak Falsafi

Computer Architecture Lab (CALCM)

- Pronounced "calcium"
- Architecture students & faculty
- "At Carnegie Mellon" means ECE + CS + ...
- 10 faculty
- Lots of very stellar students (over 30 at last count)
- Over a dozen research projects
- Seminar on Tuesdays @ 4pm



http://www.ece.cmu.edu/CALCM

© 2004 Babak Falsafi

What Are Hot Topics in Architecture? Topic #1: Parallelism

Want to make programs go fast

Execute program in parallel

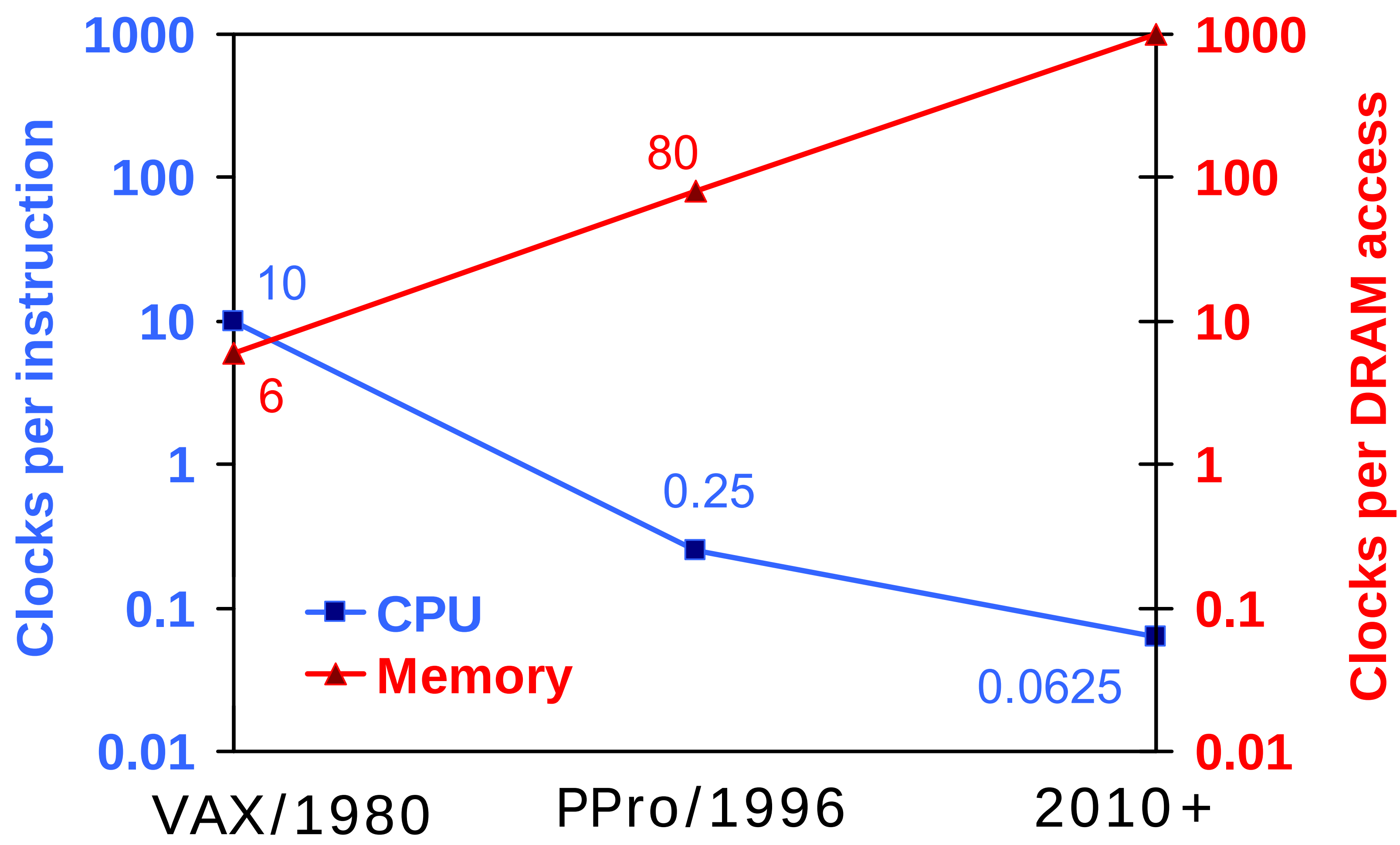
- Pipeline execution
- Superscalar → multiple pipelines
- Like mult. lunch buffets

Research:

- Independent exec?
- Program control flow?

© 2004 Babak Falsafi

Topic #2: The Memory Wall



Processor/memory performance gap will continue to increase!

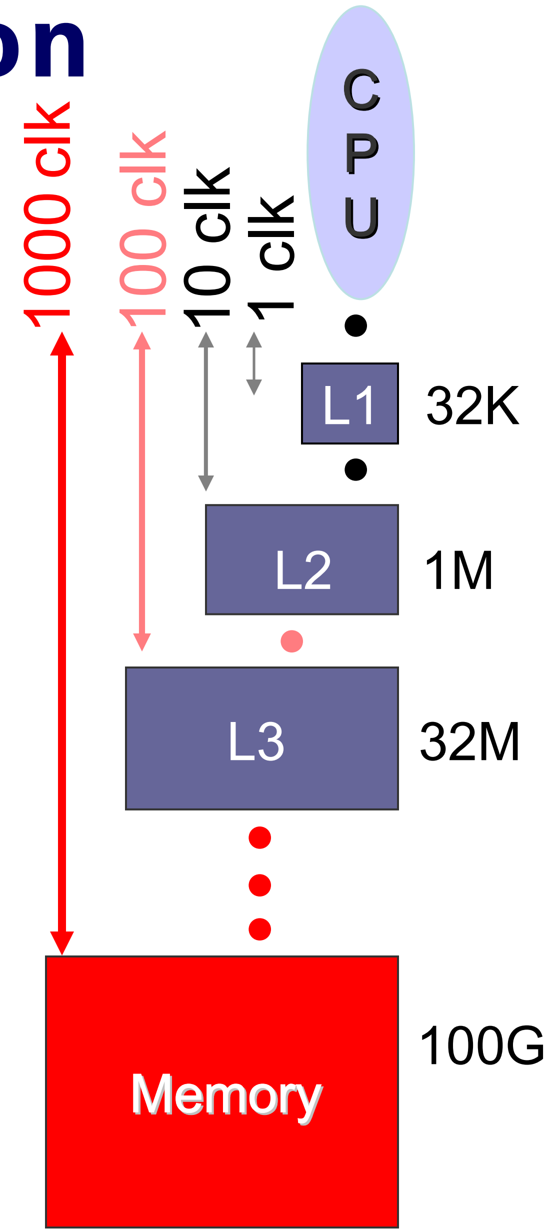
© 2004 Babak Falsafi

Today's Solution

- Hierarchy of memory levels
- Trade off capacity for speed
- Like fast short term memory
- When looking, CPU must wait

Recent research indicates

- Server running top 4 DB systems
- Half of the time, CPU is idle waiting for memory



© 2004 Babak Falsafi

Bigger Problem in Multiprocessors

Typical platforms:

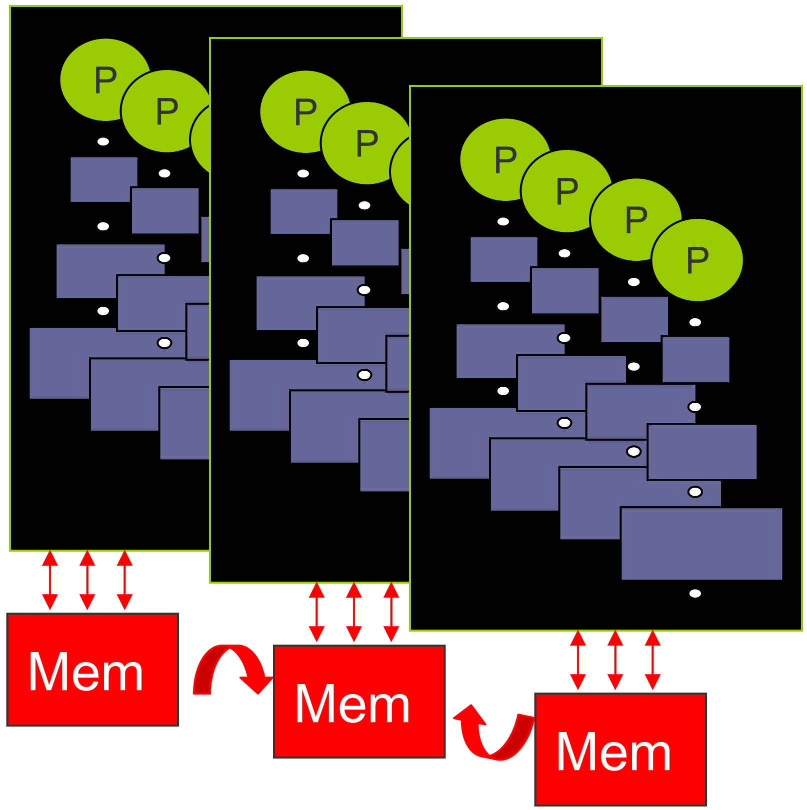
1. Chips with multiple CPUs
2. Servers with multiple chips
3. Memory shared across

Memory access:

- Traverse multiple hierarchies

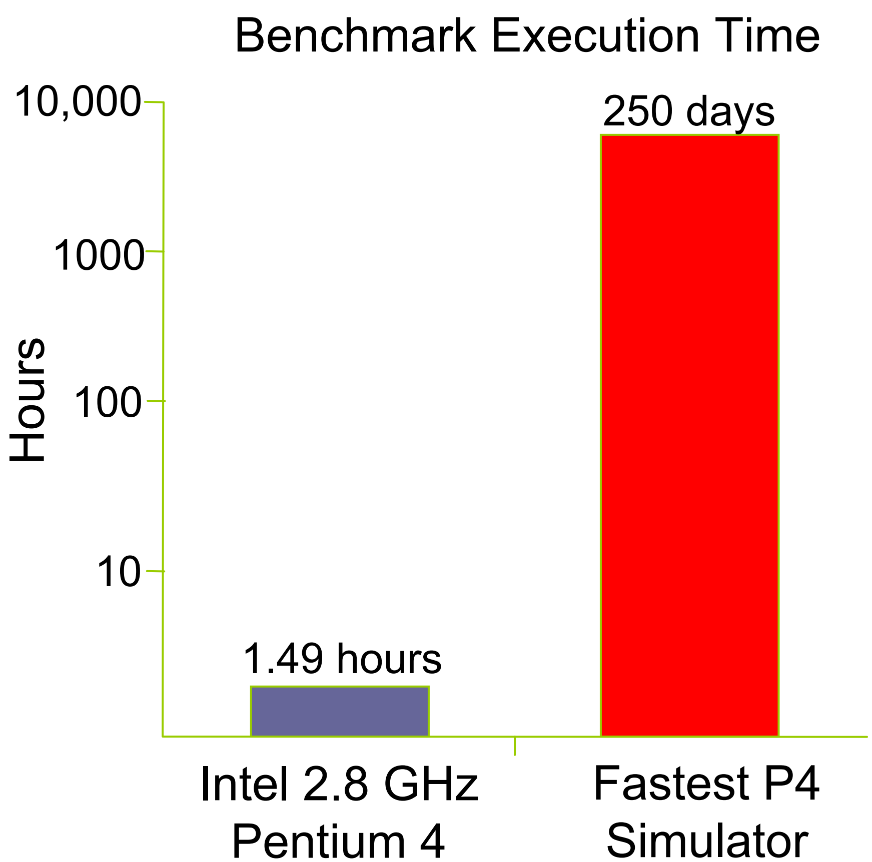
Current research:

- Track repetitive patterns
- Replay ahead of time to overlap latency



© 2004 Babak Falsafi

Topic #3: Slow Design Evaluation Tools



> 7 trillion inst. in SPEC CPU benchmarks (www.spec.org)

Current research:

- Statistical sampling
- Accurate & fast measurement

Full-benchmark simulation is not tractable

© 2004 Babak Falsafi

Topic #4: Future CMOS Digital Systems

- In 2010, we will have chips with billions and billions of tiny but lousy transistors
  - "Leak" power → need cooling
  - Lose information → need protection
  - Highly defective → cost a lot
  - Complex designs → need verification

- Research:
  - Circumvent trans. limitations

© 2004 Babak Falsafi

Reliability in Future CMOS Systems

- Highly-integrated chips
  - more error-prone
- High rate of failure in future
  - hard failure or bit flips due to cosmic rays
- High complexity designs

Reliability/availability is key in many important enterprise applications

- Downtime cost for brokerage operations: $6,450,000/h [Patterson, ROC keynote]

Current research:

- Make systems “bullet-proof” for critical apps

© 2004 Babak Falsafi

Power in Future CMOS Systems

Power consumption is hitting the roof

- Chip power soon approaching 1KW
- Buildings → Hundreds of multi-KW desktops/servers
- Battery is classic problem example
- You also run out of electricity → e.g., California

But, heat density prohibitive in high-perf. systems!

- Heat adversely affects reliability (remember AMD?)
- Some rack-mounted blades are already heat-limited

Current research:

- Reduced energy/heat with minimal perf. impact

© 2004 Babak Falsafi

Topic #5: Beyond CMOS

Molecular technology:

- Carbon nanotubes
- Act as both wires and switches
- Trillions, trillions of "transistors"

Quantum technology:

- At its very infancy

© 2004 Babak Falsafi