CS152 – Computer Architecture and Engineering

Lecture 9 – Performance

2004-09-28

Dave Patterson

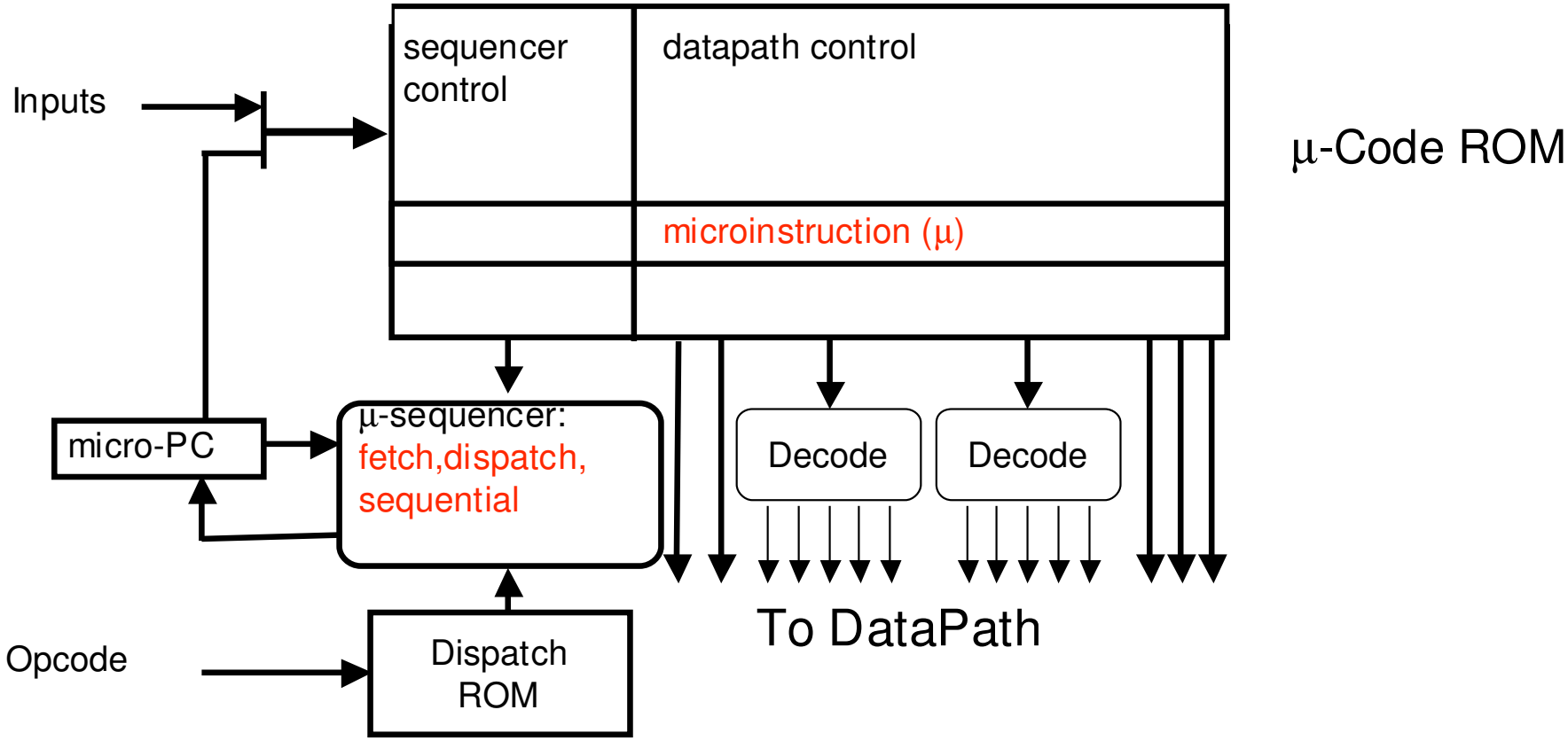
(www.cs.berkeley.edu/~patterson)

John Lazzaro

(www.cs.berkeley.edu/~lazzaro)

www-inst.eecs.berkeley.edu/~cs152/

Last Time: Microcode, Multi-Cycle



Today's Lecture - Performance

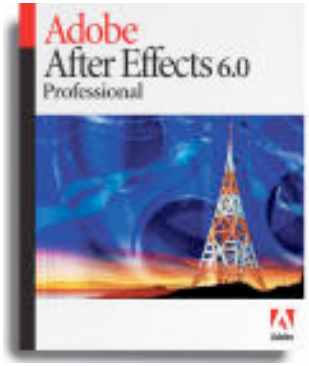
- Measurement: what, why, how
- The performance equation
- Amdahl's law
- How energy limits performance

Performance Measurement

(as seen by the customer)

Who (sensibly) upgrades CPUs often?

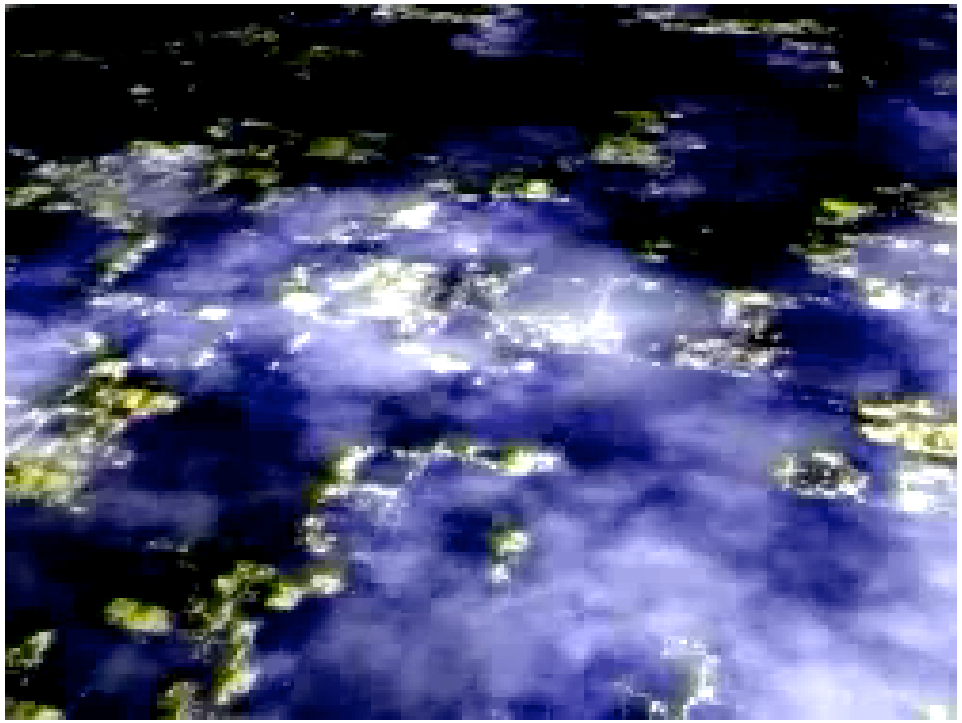
A professional who turns CPU cycles into money, and who is cycle-limited.



Artist tool: animation, video special effects.

How to decide to buy a new machine?

Measure After Effects "execution time" on a representative render "workload"



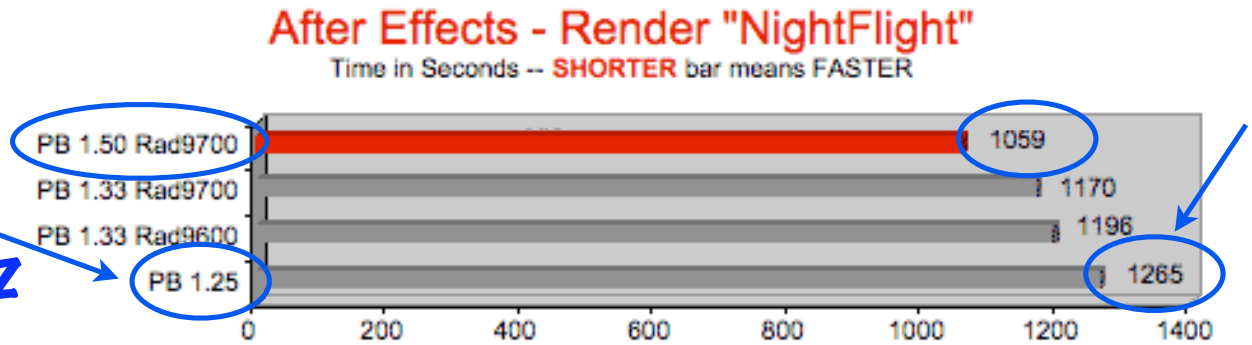
"Night flight"

City map and clouds computed "on the fly" with fractals

CPU intensive
Trivial I/O

Interpreting Execution Time

Power Book G4 1.25 GHz



Execution Time: 1265 seconds

$$\text{Performance} = \frac{1}{\text{Execution Time}} = 2.85 \text{ renders/hour}$$

1.5 GHz PB (Y) is N times faster than 1.25 GHz PB (X). N is ?

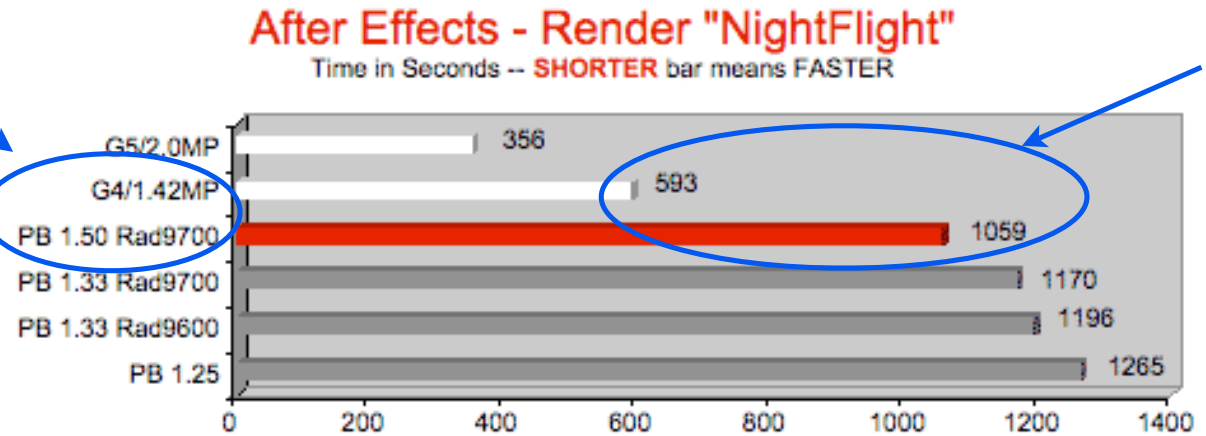
$$N = \frac{\text{Performance (Y)}}{\text{Performance (X)}} = \frac{\text{Execution Time (X)}}{\text{Execution Time (Y)}} = 1.19$$

PB 1.5 Ghz : 3.4 renders/hour. PB 1.25 : 2.85 renders/hour.
Does artist productivity really increase?

2 CPUs: Execution Time vs Throughput

Execution Time: Time for 1 job to complete

2 CPUs vs 1 CPU, otherwise similar

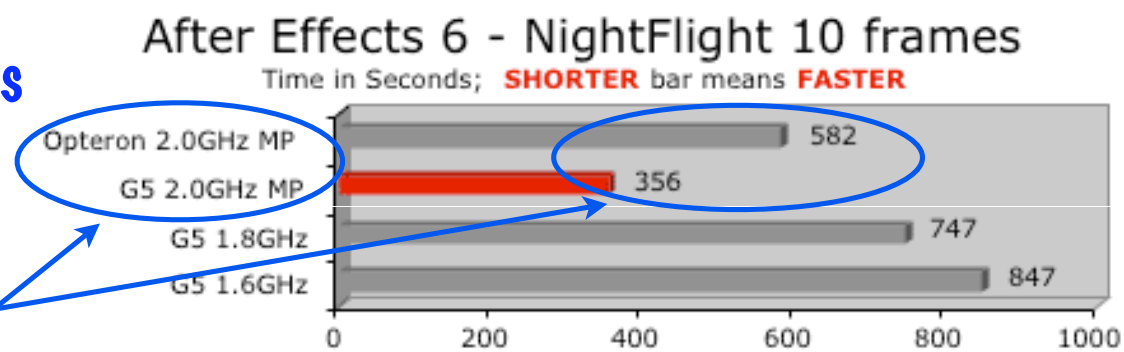


1.8x faster. What does this imply?

Throughput: # jobs/hour completed (not serialized)

Assume G5 MP execution time faster because AE does not use both Opteron CPUs.

Could G5 and Opteron have similar Throughput? Why?



Performance Measurement

(as seen by a CPU designer)

Q. Why do we care about After Effect's performance?

A. We want the CPU we are designing to run it well !

Step 1: Analyze the right measurement!

Guides CPU design

CPU Time:

Time the CPU spends running program under measurement.

How do designers use these two numbers?

How to measure CPU time?

```
% time <program name>
25.77u  0.72s  0:29.17  90.8%
```

Response Time:

Guides system design

Total time: CPU Time + time spent waiting (for disk, I/O, ...).

Administrivia - Adjust Class Time ?

We have permission to stay in this room past 12:30.

Does anyone have a class that starts 12:40?

Class time options (all “sharp” time)

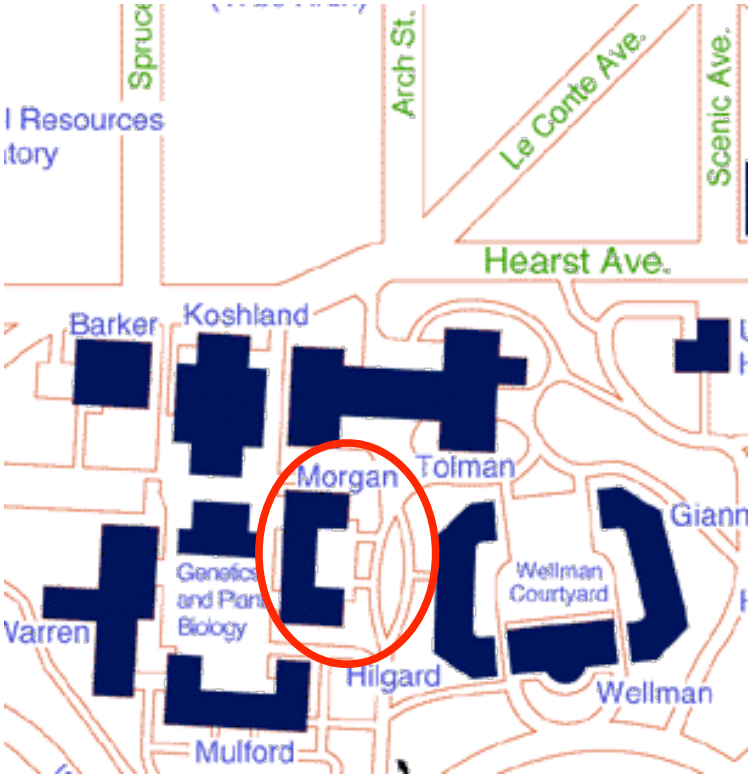
A: Lecture from 11:10 to 12:30

B: Lecture from 11:15 to 12:35

C: Lecture from 11:20 to 12:40

Administrivia - Mid-Term is Coming!

Mid-term: Tuesday 10/12, 5:30-8:30 PM, 101 Morgan. No class on Tuesday.



After exam: Pizza at LaVal's, on us!

Mid-term review session: Sunday 10/10, 7-9 PM, 306 Soda.

Administrivia - This Week's Deadlines

- Homework 2 due 9/29 (tomorrow)!
  283 Soda, in CS 152 box at 5 PM

- Lab 2 Xilinx demo on Friday 10/1
  Lab 2 due Monday 10/4, 11:59 PM

  On Tuesday 10/5, onto the Pipelining Lab!

CPU time: Proportional to Instruction Count

Q. Once ISA is set, who can influence instruction count?

A. Compiler writer, application developer.

Q. Static count? (lines of program printout) Or dynamic count? (trace of execution)

A. Dynamic.

$$\frac{\text{CPU time}}{\text{Program}} \propto \frac{\text{Machine Instructions}}{\text{Program}}$$

Rationale: Every additional instruction you execute takes time.

Q. What type of computer architect influences the number of instructions a given program needs?

A. Instruction set architect.

CPU time: Proportional to Clock Period

Q. How can architects (not technologists) reduce clock period?

A. Shorten the machine critical path.

Q. What ultimately limits an architect's ability to reduce clock period ?

A. Clock-to-Q, setup times.

$$\frac{\text{Time}}{\text{Program}} \propto \frac{\text{Time}}{\text{One Clock Period}}$$

Rationale:

We measure each instruction's execution time in "number of cycles". By shortening the period for each cycle, we shorten execution time.

Completing the performance equation

What factors make the CPI for a program differ from the underlying CPI of a CPU implementation?

- Cache behavior varies.
- **Instruction mix varies**
- Branch prediction varies.

$$\frac{\text{Seconds}}{\text{Program}} = \frac{\text{Instructions}}{\text{Program}} \; \frac{\text{Cycles}}{\text{Instruction}} \; \frac{\text{Seconds}}{\text{Cycle}}$$

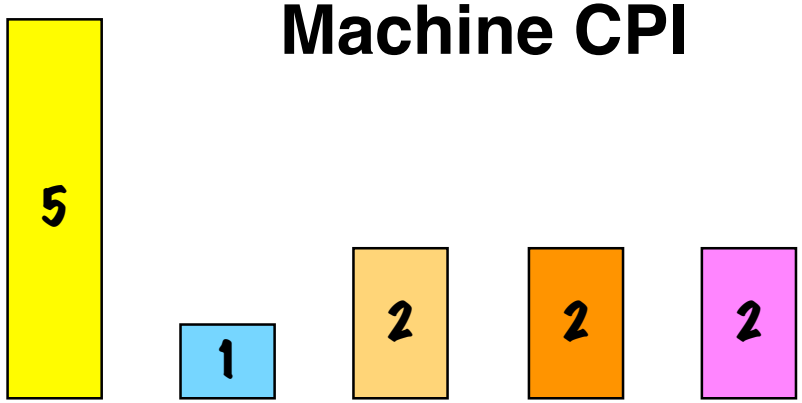
We need all three terms, and only these terms, to compute CPU Time!

"CPI" -- The Average Number of Clock Cycles Per Instruction → For the Program ←

When is it OK to compare clock rates?

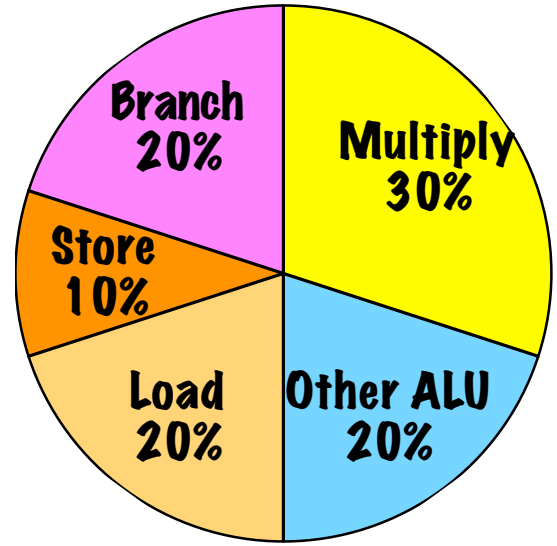
CPI as an analytical tool to guide design

Machine CPI



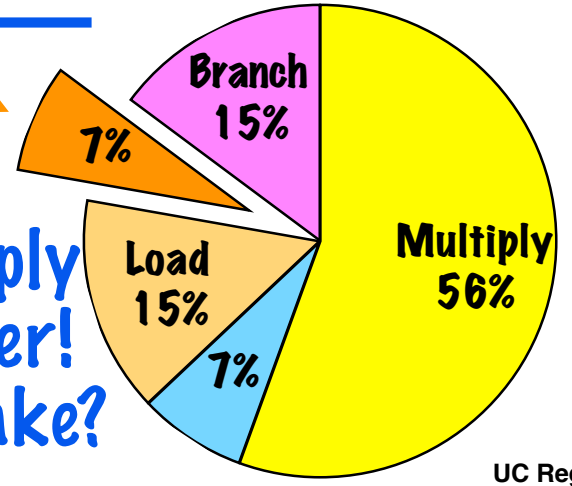
Multiply Other ALU Load Store Branch

Program Instruction Mix



$$\frac{5 \times 30 + 1 \times 20 + 2 \times 20 + 2 \times 10 + 2 \times 20}{100}$$

= 2.7 cycles/instruction

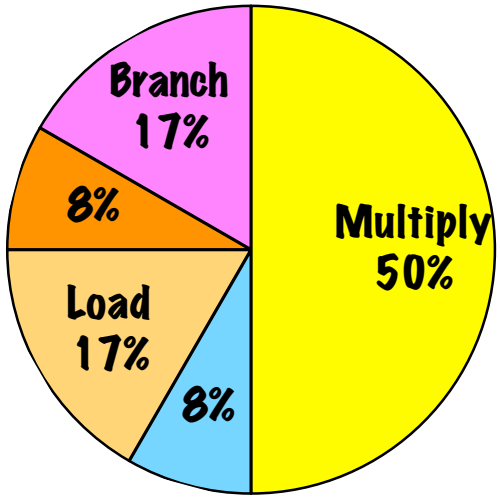


Where program spends its time

Q. We lower machine multiply CPI, but program runs slower! What mistake(s) did we make?

Amdahl's Law (of Diminishing Returns)

Where program spends its time



If enhancement "E" speeds up multiply, but other instructions are unchanged, what is the maximum speedup S?

$$S_{max} = \frac{1}{1 - (\% \text{ affected} / 100\ \%)} = \frac{1}{1 - (50/100)} = 2$$

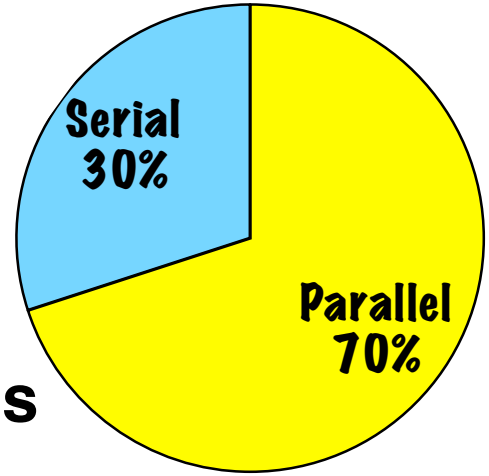
Attributed to Gene Amdahl -- "Amdahl's Law"

What is the lesson of Amdahl's Law?

Must enhance computers in a balanced way!

Peer Instruction: Amdahl's Law

Program
We
Wish
To Run
On N CPUs



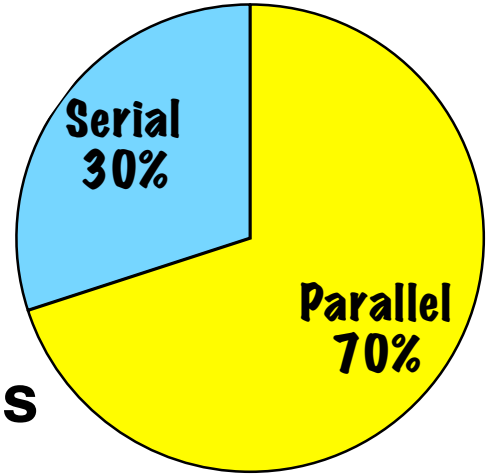
The program spends 30% of its time running code that can not be recoded to run in parallel.

Compute speedup for N = 2, 3, 4, 5, and ∞.

CPUs	2	3	4	5	∞
Speedup					

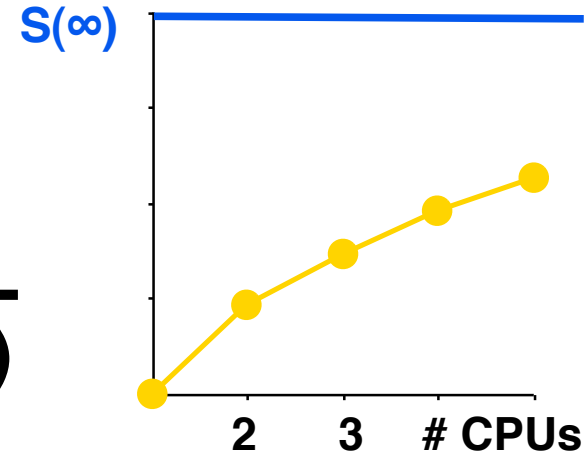
Peer Instruction: Amdahl's Law

Program We Wish To Run On N CPUs



The program spends 30% of its time in serial code.

Compute speedup for N = 2, 3, 4, 5, and ∞.



$$S = \frac{1}{1 - (30\ \% + (70\% / N)\ ) / 100\ \%)}$$

CPUs	2	3	4	5	∞
Speedup	1.54	1.85	2.1	2.3	3.3

Final thoughts: Performance Equation

$$\frac{\text{Seconds}}{\text{Program}} = \frac{\text{Instructions}}{\text{Program}} \quad \frac{\text{Cycles}}{\text{Instruction}} \quad \frac{\text{Seconds}}{\text{Cycle}}$$

Goal is to optimize execution time, not individual equation terms.

Machines are optimized with respect to program workloads.

The CPI of the program. Reflects the program's instruction mix.

Clock period. Optimize jointly with machine CPI.

1 Joule of energy is dissipated by a 1 Amp current flowing through a 1 Ohm resistor for 1 second. Also, 1 Watt for 1 second.

1 Watt: 1 Amp flowing through 1 Ohm.

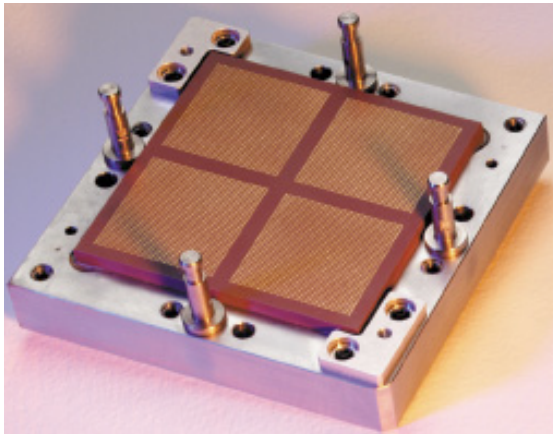
Energy and Performance

1 Joule = 0.24 calories.
1 calorie raises 1 gram of water 1 °C

Sad fact: computers turn electrical energy into heat. Computation is a byproduct.

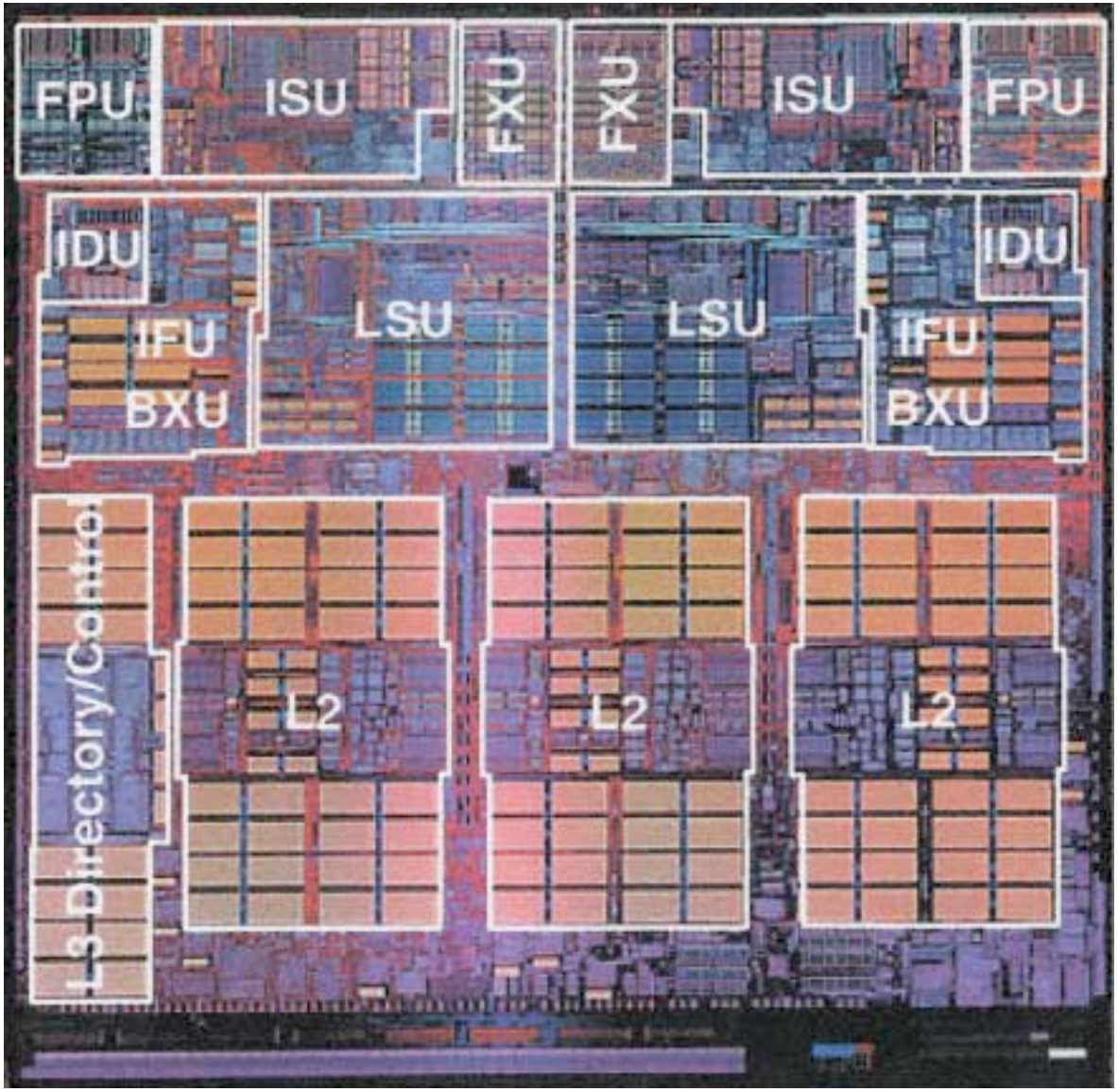
Air or water carries heat away, or chip melts.

IBM Power 4: How does die heat up?

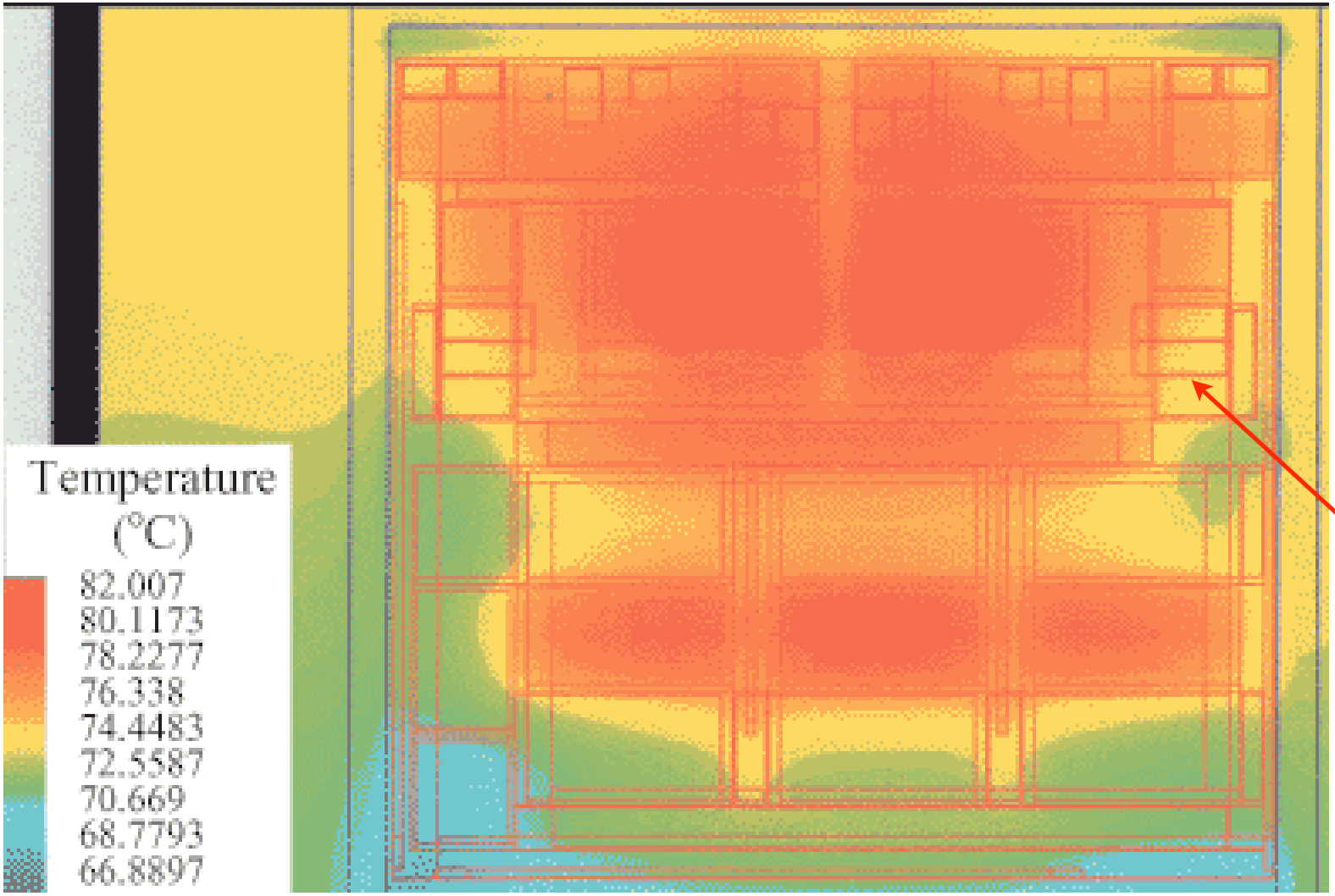


4 dies on a multi-chip module

2 CPUs per die



IBM Power 4: Dissipating 115 Watts



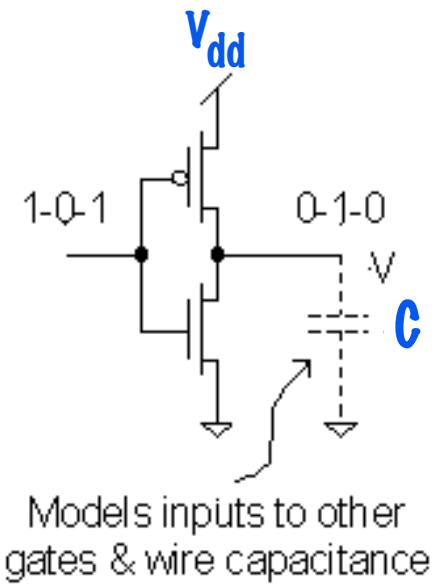
Hot spots

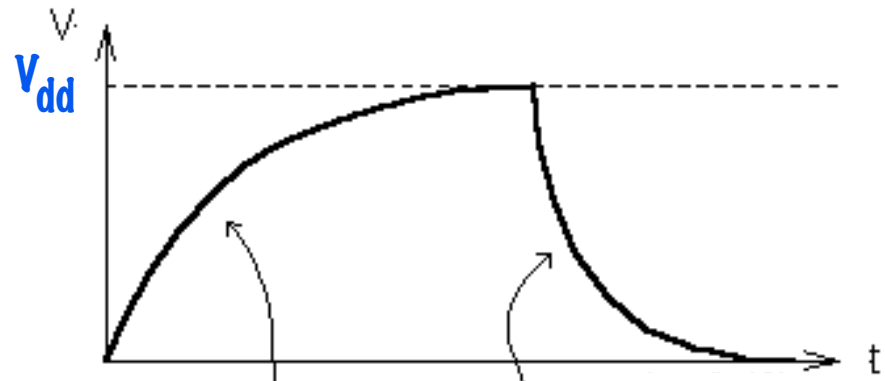
Fixed point units

Cache logic

Switching energy: Fundamental Physics

Every logic transition dissipates energy.





$$E_{0\to1} = \frac{1}{2} C V_{dd}^2 \qquad E_{1\to0} = \frac{1}{2} C V_{dd}^2$$

Strong result: Independent of technology.

How can we limit switching energy?

State-of-the-art CPUs (90 nm):

Switching energy is 70% of total energy.

Remainder: at 90nm, "switches" are "dimmers"!

"leakage" currents 65nm: 50/50!

Conclusions

- Customers: measure to buy
  Architects: measure for design
- Tools: Performance Equation, CPI
- Amdahl's Law's lesson: Balance
- Energy: $E_{0\to1} = \frac{1}{2} C V_{dd}^2$ $E_{1\to0} = \frac{1}{2} C V_{dd}^2$