Lecture 8: More performance

Friday, January 25, 2019 10:40 AM

Outline

- More Iron law examples
- Comparing performance of systems
- Intro to pipelining

$\boxed{\text{System A:}}$ CPI → 5 @ 500 MHz
System B: CPI → 3 @ 200 MHz } application: 1 million instructions

$$time_A = \frac{sec}{cycle} \cdot \frac{cycle}{inst.} \cdot \#\,inst. = \frac{1}{500\,10^6} \cdot 5 \quad 1 \cdot 10^6 = 10\ ms = \frac{1}{100}\ s$$

$$time_B = \frac{1}{200\,10^6} \cdot 3 \cdot 10^6 = \frac{3}{200}\ s = 15\ ms$$

$t_A < t_B$

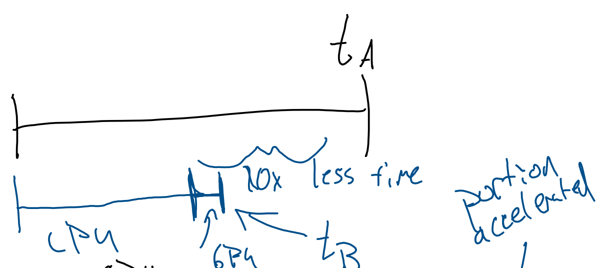
$\frac{t_A}{t_B}$ → time for A is $2/3$ time for B

$\frac{t_B}{t_A}$ → 1.5 x t_A is 1.5 times faster than t_B
System A System B

or system A provides a 1.5x speedup over System B

$$Speedup = \frac{\text{time for old}}{\text{time for new}}$$

System A → CPU 10s
System B → CPU + GPU



50 % of app sees 10x speedup on GPU compared to CPU

Speedup of System B over System A?

Speedup = 10x?
< 10x?

$$t_{new} = t_{old} \cdot .5 + \frac{.5 \cdot t_{old}}{10}$$
$$= 10 \cdot .5 + \frac{.5 \cdot 10}{10}$$
$$= 5.5s$$

speedup

$$\frac{t_{old}}{t_{new}} = \frac{10s}{5.5s}$$

$\boxed{\text{speedup} = 1.81\ x}$

Amdahl's Law

Amdahl's ...

$$\text{Speedup} = \frac{1}{(1-F_e) + \frac{F_e}{S_e}}$$

fraction enhanced

speedup of enhancement

$$\text{Speedup}_n = \frac{1}{1-p+\frac{p}{n}}$$

$n \leftarrow$ # of procs

parallel portion

$$\text{speedup} = \frac{t_{old}}{t_{new}} = \frac{1}{1-f_e+\frac{f_e}{S_e}}$$

Parallel programs

part of the program is inherently serial

parallel fraction < 90%

$$t_{new} = t_{old} \cdot (1-F_e) + t_{old} \cdot \frac{F_e}{S_e}$$

90% parallel speedup w/ 2 procs over 1 proc?

$$\text{Speedup} = \frac{1}{1-.9+\frac{.9}{2}} = \frac{1}{.1+.45} = \frac{1}{.55} = 1.82x$$

$$= \frac{1}{1-.9+\frac{.9}{4}} = \frac{1}{.1+.225} = 3.08x$$

most speedup? parallel portion in 0 time $\rightarrow$ $\frac{1}{1-.9} = \frac{1}{.1} = \boxed{10x}$

Gustafson's Law

$\hookrightarrow$ speedup with respect to data size growing

	Baseline system	System A	system B
inst. mix	30% lds	1x	1x
	20% st.	1x	1x
	10% div	5x speedup	1x
	40% ALU	1x	1.5x speedup

Speedup A: Amdahl's law $\rightarrow$ $\frac{1}{1-.1+\frac{.1}{5}} = \frac{1}{.92} = 1.09x$

$\boxed{\text{Speedup B:}}$ " " $\rightarrow$ $\frac{1}{1-.4+\frac{.4}{1.5}} = \frac{1}{.6+\frac{.8}{3}} = 1.15x$

$S_B > S_A$

$$\frac{S_B}{S_A} = \frac{1.15x}{1.09x} = 1.055 \rightarrow$$ speedup of B over A

Perf of single-cycle CPU

latency + physics

↳ latency is mostly fixed

↳ due to no more Dennard scaling + Moore's Law

How to incr. perf?

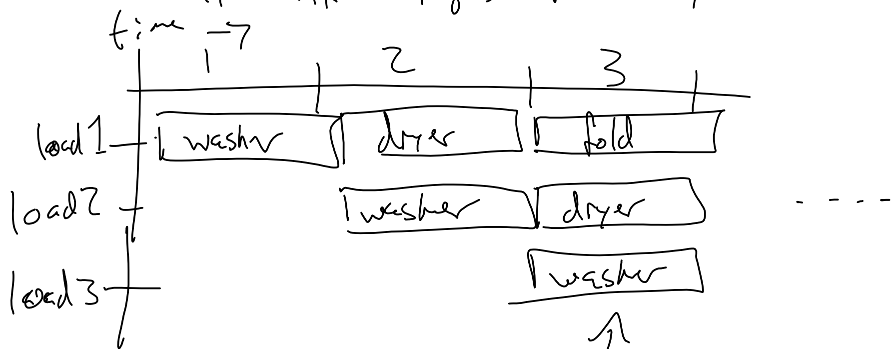
<u>parallelism</u> ☆ → incr bandwidth $\frac{\#\text{ things done}}{\text{cycle}}$

↳ Laundry

washer ~ 1hr + dryer ~ 1h + fold ~1h

1 load takes 3 hrs

Overlap diff. stages of laundry



3 loads in parallel

latency for a load of laundry?

3 hrs.

what's the throughput or bandwidth?

1 load / hour