Evaluating Performance with Speedup

Chris Rossbach and Calvin Lin

Sequential Algorithm Analysis

We use algorithm analysis to compare algorithms

Algorithm analysis counts abstract units of time

e.g. merge sort: $O(n \log n)$ time

e.g. bubble sort: $O(n^2)$ time

Why not just run a program and report its execution time?

Limitations of Execution Time

Execution time depends on many factors

- Problem size
- Clock speed
- Processor architecture
- Language and compiler
- . . .

Algorithm analysis is insensitive to these details

Algorithm analysis is technology-independent

Parallel Speedup

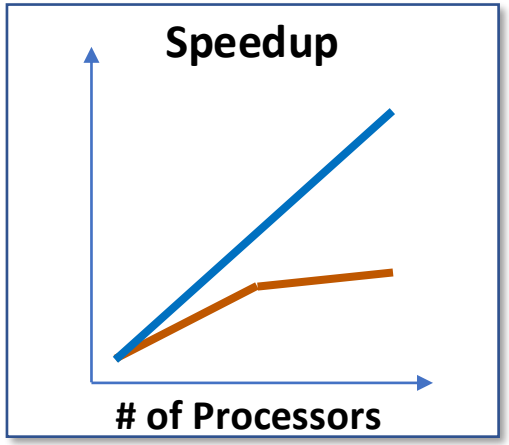
$$\text{Speedup} = \frac{\text{Serial Time}}{\text{Parallel Time}}$$

One way to cheat is to use a bad algorithm for the Serial Time.

Are there other issues with speedup as a metric?

Expectations

Is this good speedup?



What speedup should we expect?

Some problems admit solutions that are embarrassingly parallel

Other problems admit solutions for which it is much more difficult to extract parallel performance

Dependence on Problem Size

Even for a fixed problem, speedup typically depends on the problem size

How do we know this?

Gustafson's Law

The $n_{1/2}$ Metric [Hockney 77]:

What problem size, n, is needed to run a given program on a given machine to achieve parallel efficiency of $^1/_2$?

(Recall that Efficiency = Speedup / P)

Low latency communication $\Rightarrow$ smaller value of n

For a fixed problem size, lower latency $\Rightarrow$ greater speedup

Back to Speedup

What happens to speedup if we move to a slower processor?

The relative latency of communication decreases, so speedup increases

What happens if we use an inferior compiler?

What if we turn off optimization flags?

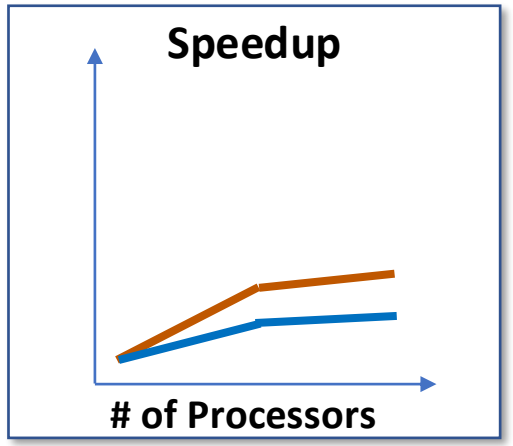
What if we use single-precision arithmetic?

Speedup is technology-dependent

Relative Performance

With all of these factors, how should we interpret speedup?

Compare your solution to a state-of-the-art solution using the same methodology



Summary

Speedup is technology-dependent

Be careful when reporting and interpreting speedup

It’s difficult to interpret speedup in isolation