HW9.7. Amdahl's Law

Amdahl's law allows for the calculation of the maximum expected speedup of a system when only part of the system can be optimized/parallelized. It presents the idea that the overall speedup of a program will be limited by the part that cannot be optimized. The formula demonstrating this concept is shown below:

$$S = \frac{1}{(1-P) + (\frac{P}{N})}$$

S is the overall speedup of the program. There is a speedup if this is greater than 1.

P is the fraction of the program that can be optimized/parallelized (this is less than 1 if not everything can be optimized). $(1-P)$ would be the fraction of the program that cannot be optimized.

N is the optimization/parallelization factor that can only be applied to the fraction **P**. This should be greater than 1 if you are doing optimizations.

Q1.1: Your friend is analyzing a website and notices that database operations take up 2/3 of the webpage's total loading time. Database operations previously took 300ms per request, and your friend reduced this down to 150ms. What is the speedup observed by the user?

1.50 ? ✓ 100%

Q2: We have a system to which we can instantaneously add and remove cores – adding more cores never leads to slowdown from things like false sharing, thread overhead, context switching, etc. When the program foo() is executed to completion with a single core in the system, it completes in 20 minutes. When foo() is run with a total of four cores in the system, it completes in 8 minutes.

Q2.1: If 100% of foo() was parallelized, with 4 cores it would take 20/4 = 5 minutes. Since it instead takes 8 minutes, what fraction of foo() was parallelized?

0.8 ? ✓ 100%

Q2.2: Since we have this magic system and we *really* need foo to run, we decide to buy all the computing resources on the planet, effectively adding infinitely many cores to our system. How many minutes does foo take to execute now?

4.0 ? ✓ 100%

Try a new variant

Correct answer

Amdahl's law allows for the calculation of the maximum expected speedup of a system when only part of the system can be optimized/parallelized. It presents the idea that the overall speedup of a program will be limited by the part that cannot be optimized. The formula demonstrating this concept is shown below:

$$S = \frac{1}{(1-P) + (\frac{P}{N})}$$

S is the overall speedup of the program. There is a speedup if this is greater than 1.

P is the fraction of the program that can be optimized/parallelized (this is less than 1 if not everything can be optimized). $(1-P)$ would be the fraction of the program that cannot be optimized.

N is the optimization/parallelization factor that can only be applied to the fraction **P**. This should be greater than 1 if you are doing optimizations.

Q1.1: Your friend is analyzing a website and notices that database operations take up 2/3 of the webpage's total loading time. Database operations previously took 300ms per request, and your friend reduced this down to 150ms. What is the speedup observed by the user?
`1.50`

1: One option is to use the formula (described in the answer to question 2), but it's generally better to have a logical reasoning that allows you to reach the same result. One easy way to do this is to try and determine the speedup of a fixed number of operations. Let's say that we have 300 seconds worth of operations (300 was chosen arbitrarily to make the math end up with nice round numbers; any number of seconds can be chosen, and we would get the same speedup value at the end). Then 200 seconds would be spent on database operations. We were able to halve that time, so on our new system, it takes 100 seconds for database operations, or 200 seconds total. Thus, we went from 300 to 200 seconds, for a total speedup of 1.5x.

Q2: We have a system to which we can instantaneously add and remove cores – adding more cores never leads to slowdown from things like false sharing, thread overhead, context switching, etc. When the program foo() is executed to completion with a single core in the system, it completes in 20 minutes. When foo() is run with a total of four cores in the system, it completes in 8 minutes.

Q2.1: If 100% of foo() was parallelized, with 4 cores it would take 20/4 = 5 minutes. Since it instead takes 8 minutes, what fraction of foo() was parallelized?
`0.800`

2.1: Using the formula

$$S = \frac{1}{(1-P) + (\frac{P}{N})}$$

with $S = 20/8$, $N = 4$, and solving for P yields $P = 4/5$.

Q2.2: Since we have this magic system and we *really* need foo to run, we decide to buy all the computing resources on the planet, effectively adding infinitely many cores to our system. How many minutes does foo take to execute now?
`4.00`

2.2: For an infinite number of cores, we get

$$\lim_{N \to \infty} \frac{1}{(1-0.8) + (\frac{0.8}{N})} = \frac{1}{(0.2)} = 5$$

for our speedup. Since our original took 20 minutes, our new runtime would be $20/5 = 4$ minutes.

Submitted answer **correct: 100%**
Submitted at 2022-11-20 05:02:43 (PST)

hide

Amdahl's law allows for the calculation of the maximum expected speedup of a system when only part of the system can be optimized/parallelized. It presents the idea that the overall speedup of a program will be limited by the part that cannot be optimized. The formula demonstrating this concept is shown below:

$$S = \frac{1}{(1-P) + (\frac{P}{N})}$$

S is the overall speedup of the program. There is a speedup if this is greater than 1.

P is the fraction of the program that can be optimized/parallelized (this is less than 1 if not everything can be optimized). $(1-P)$ would be the fraction of the program that cannot be optimized.

N is the optimization/parallelization factor that can only be applied to the fraction **P**. This should be greater than 1 if you are doing optimizations.

Q1.1: Your friend is analyzing a website and notices that database operations take up 2/3 of the webpage's total loading time. Database operations previously took 300ms per request, and your friend reduced this down to 150ms. What is the speedup observed by the user?

`1.5` ? ✓ 100%

Q2: We have a system to which we can instantaneously add and remove cores – adding more cores never leads to slowdown from things like false sharing, thread overhead, context switching, etc. When the program foo() is executed to completion with a single core in the system, it completes in 20 minutes. When foo() is run with a total of four cores in the system, it completes in 8 minutes.

Q2.1: If 100% of foo() was parallelized, with 4 cores it would take 20/4 = 5 minutes. Since it instead takes 8 minutes, what fraction of foo() was parallelized?

`0.8` ? ✓ 100%

Q2.2: Since we have this magic system and we *really* need foo to run, we decide to buy all the computing resources on the planet, effectively adding infinitely many cores to our system. How many minutes does foo take to execute now?

4 ? ✓ 100%