livelessons
video instruction from technology experts

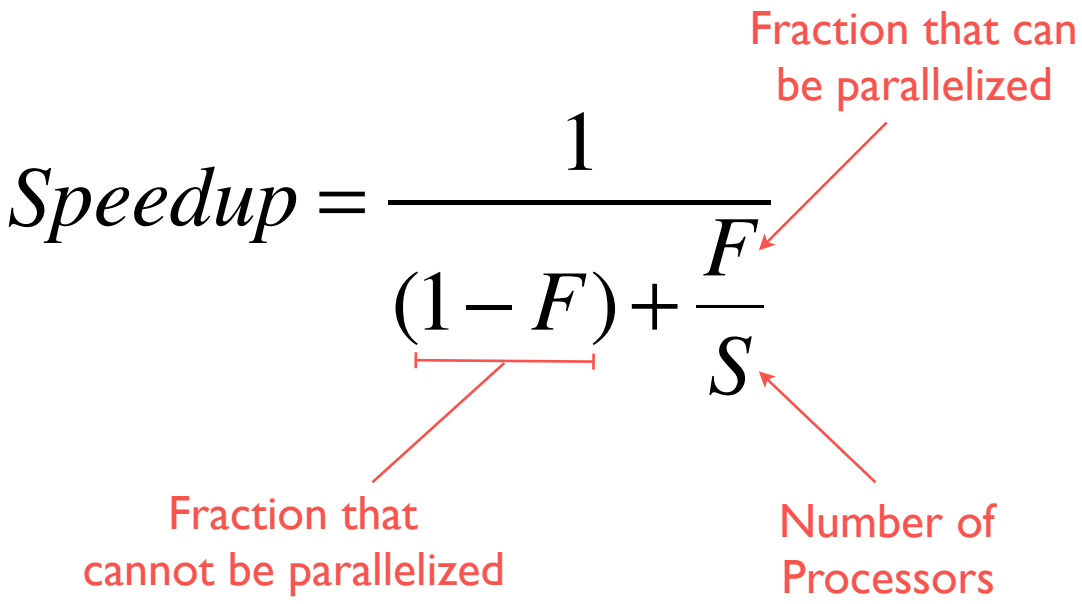
Lesson 4: Spark Internals

4.11 Making Spark Fly: Parallelism

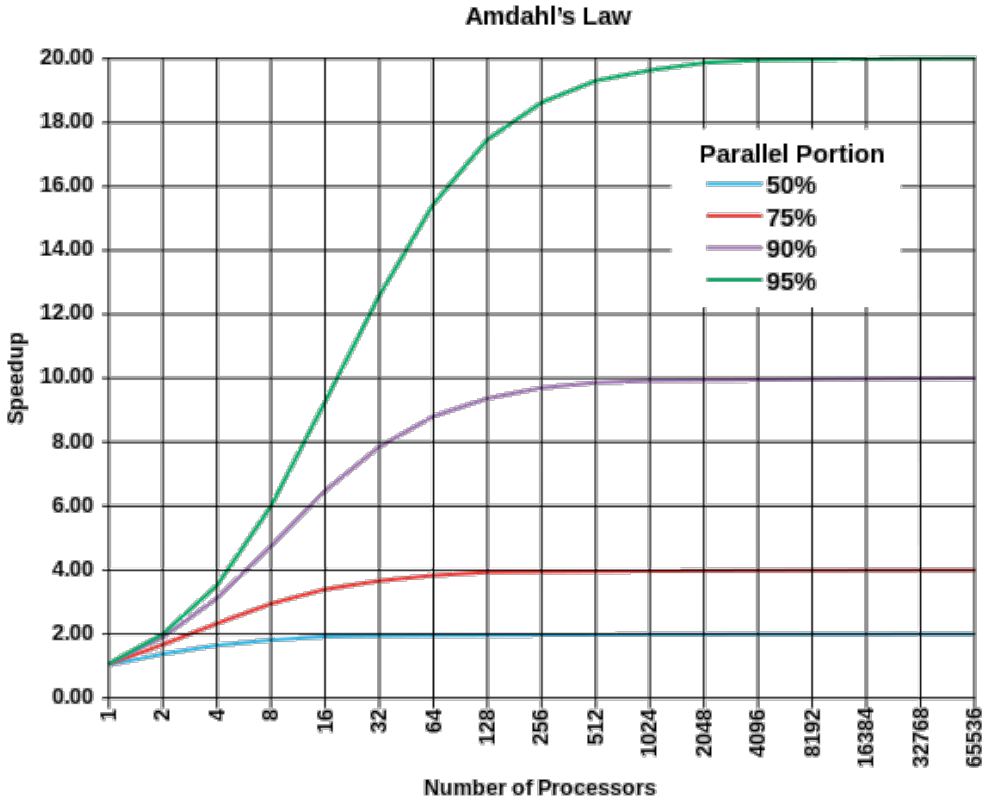
Parallelism

- Too much: `coalesce()`
  - Many tasks complete too quickly, unnecessary coordination/communication overhead
  - idle tasks waiting to get scheduled
  - < 100ms per task
- Too little: `repartition()`
  - Cannot leverage concurrency/parallelism
  - Susceptible to data skew
  - ~2x number of cores in cluster

Amdahl's Law



Amdahl's Law



Amdahl's Law

20x improvement
25% of the time

Fraction that can be parallelized

$$Speedup = \frac{1}{(1-0.25)+\frac{0.25}{20}}$$

Fraction that cannot be parallelized

Number of Processors

Amdahl's Law

20x improvement
25% of the time

Only 1.3x Speedup!

$$Speedup = \frac{1}{0.75 + \frac{0.25}{20}} = 1.31$$

Even with 20 processors

Moral?

Amdahl's Law

It doesn't how many machines you throw at a problem... at some point it is just not going to run any faster