Parallel Programming

Basic Concepts in Parallelism

Expressing Parallelism

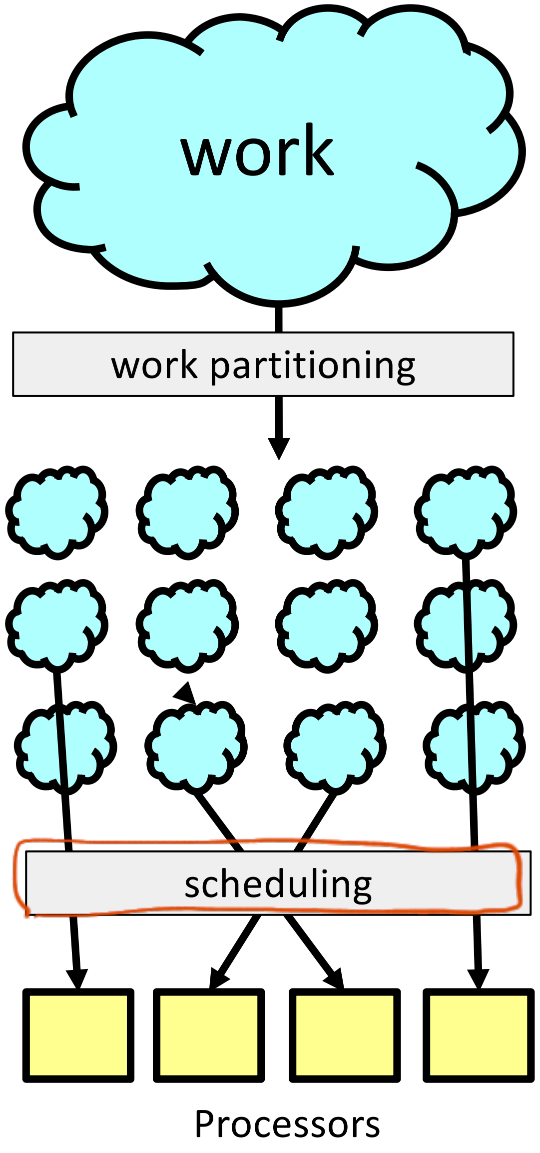
- Work partitioning
  - Split up work of a single program into **parallel tasks**

- Can be done:
  - Explicitly / Manually (**task/thread parallelism**)
    - User explicitly expresses tasks/threads
  - Implicit parallelism:
    - Done automatically by the system (e.g., in **data parallelism**)
    - User expresses an operation and the system does the rest

Work Partitioning & Scheduling

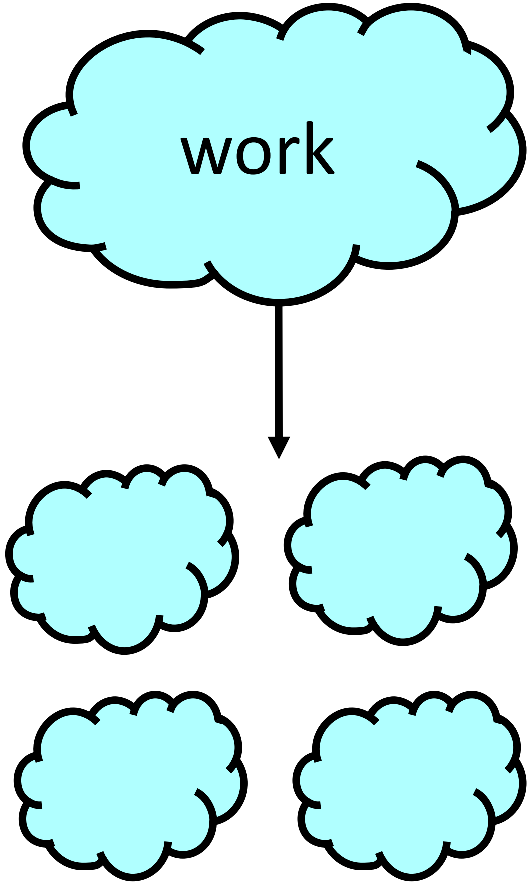
- **work partitioning**
  - **split up** work into **parallel tasks/threads**
  - (done by user)
  - A task is a unit of work
  - also called: **task/thread decomposition**
- **scheduling**
  - assign tasks to processors
  - (typically done by the system)
  - goal: <u>full utilization</u>

  (no processor is ever idle)

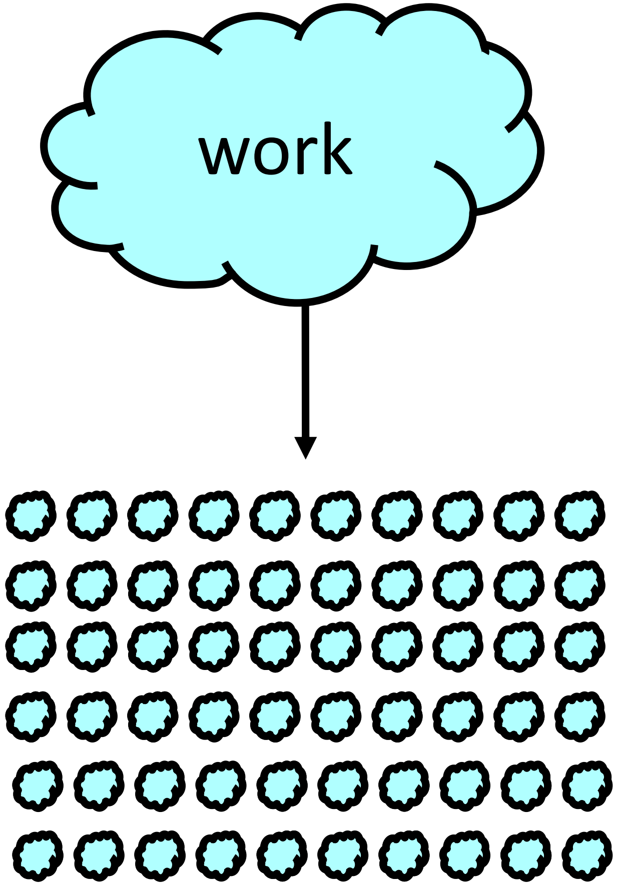


# of chunks should be larger than the # of processors

Task/Thread Granularity

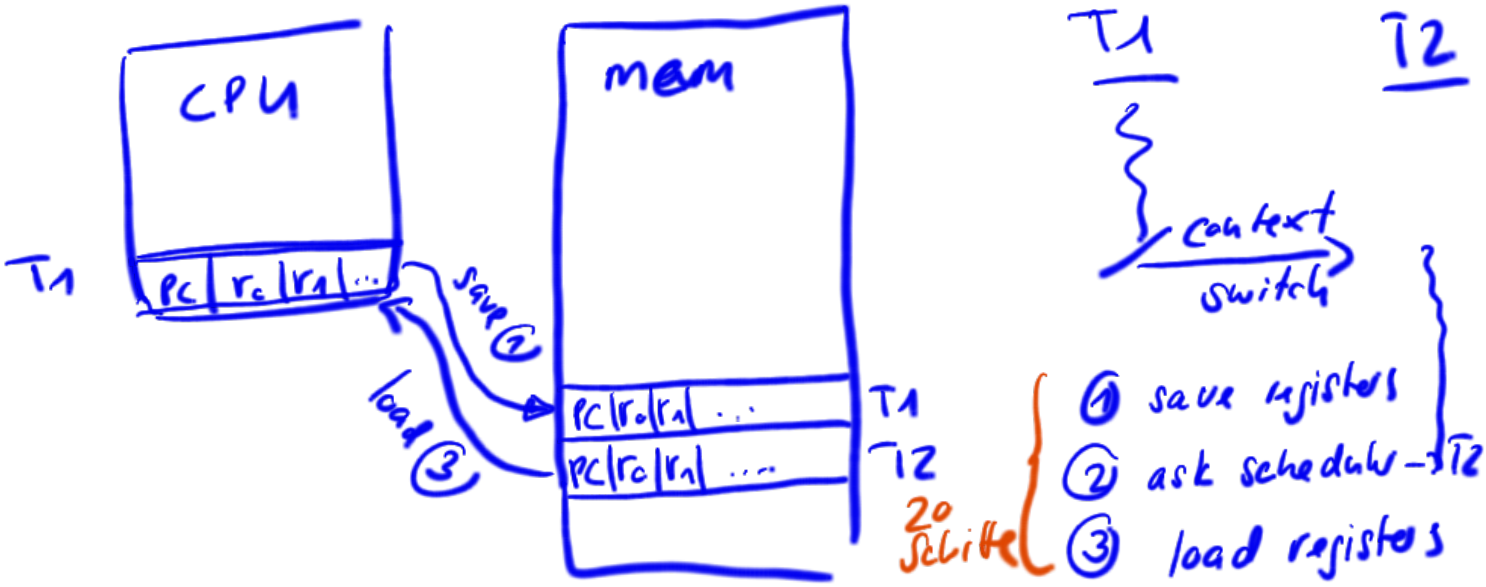


Coarse granularity



Fine granularity

Thread scheduling overhead



Coarse vs Fine granularity

- **Fine granularity:**
  - more portable

    (can be executed in machines with more processors)
  - better for scheduling
  - but: if scheduling overhead is comparable to a single task → overhead dominates

Task granularity guidelines

- As small as possible
- but, significantly bigger than scheduling overhead
  - system designers strive to make overheads small

Scalability

An overloaded concept: e.g., how well a system reacts to increased load, for example, clients in a server

In parallel programming:

- speedup when we increase processors
- what will happen if processors → ∞
- a program scales linearly → linear speedup

Parallel Performance

Sequential execution time: $\mathbf{T_1}$

Execution time $\mathbf{T_p}$ on $\mathbf{p}$ CPUs

- $\mathbf{T_p = T_1 / p}$ (perfection)
- $\mathbf{T_p > T_1 / p}$ (performance loss, what normally happens)
- $\mathbf{T_p < T_1 / p}$ (sorcery!)

(parallel) Speedup

(parallel) speedup S_p on **p** CPUs:

$$S_p = T_1 / T_p$$

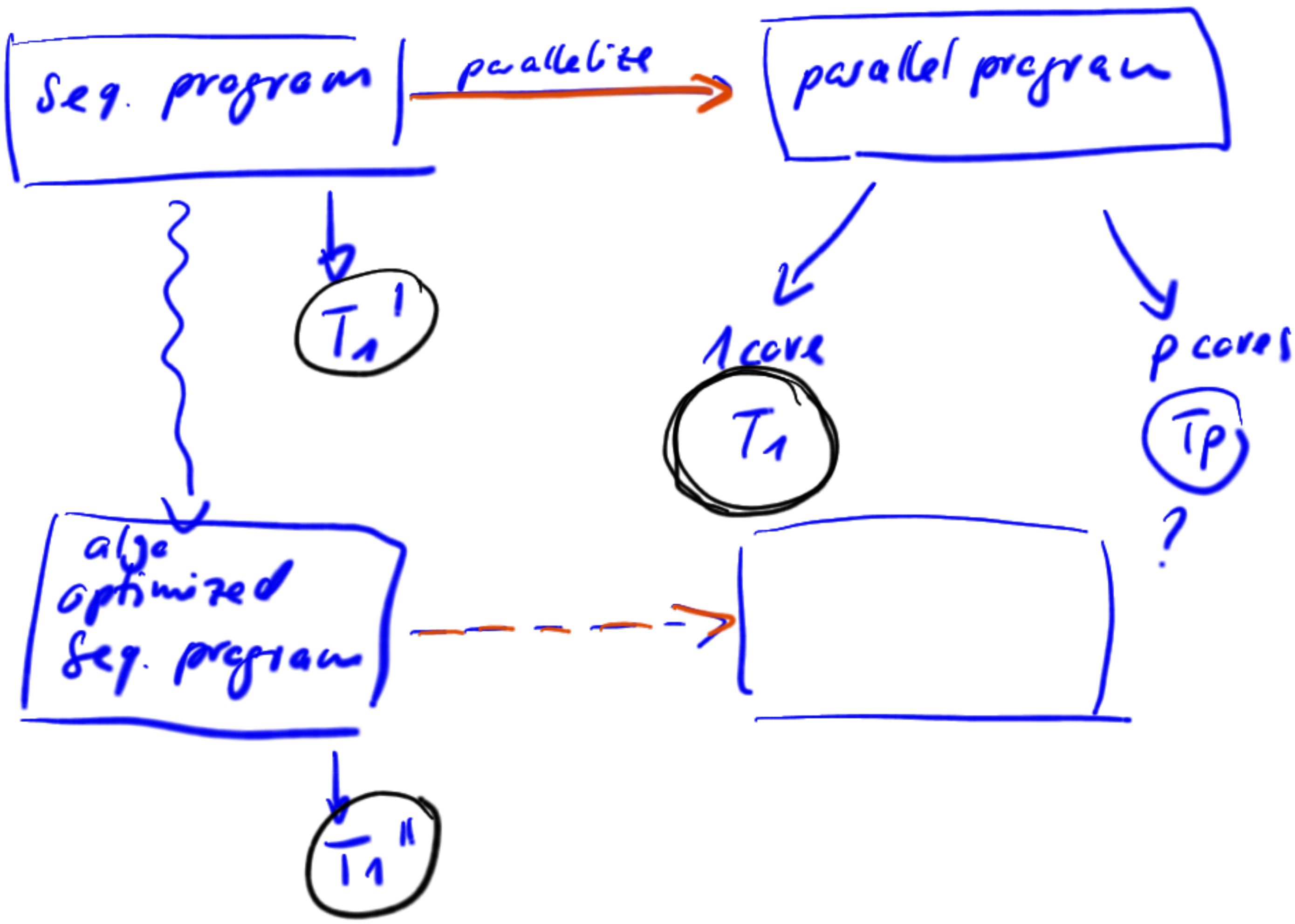
- $S_p = p$ → linear speedup (perfection)
- $S_p < p$ → sub-linear speedup (performance loss)
- $S_p > p$ → super-linear speedup (sorcery!) $S_p = 100$
- **Efficiency: S_p / p** $= 1$ 100%

$p = 1000$ cores

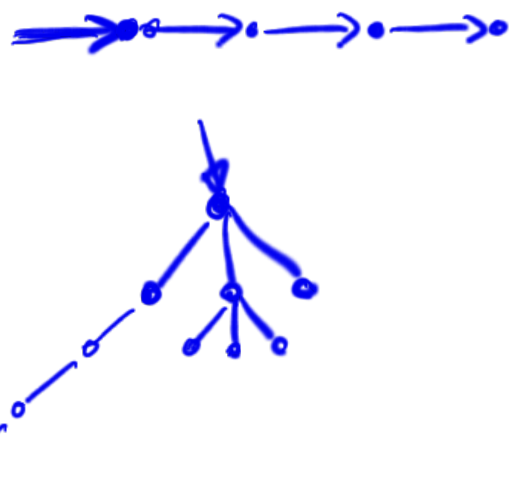
$E = \frac{100}{1000} = 0.1$ 10%

what is T_1?

$T_1'' \ll T_1' < T_1$



$$S_p = \frac{T_1}{T_p}$$



$S_p \leq 1$

$S_p \gg 1$

Absolute versus Relative Speed-up

Relative speedup (Efficiency):

relative improvement from using P execution units.

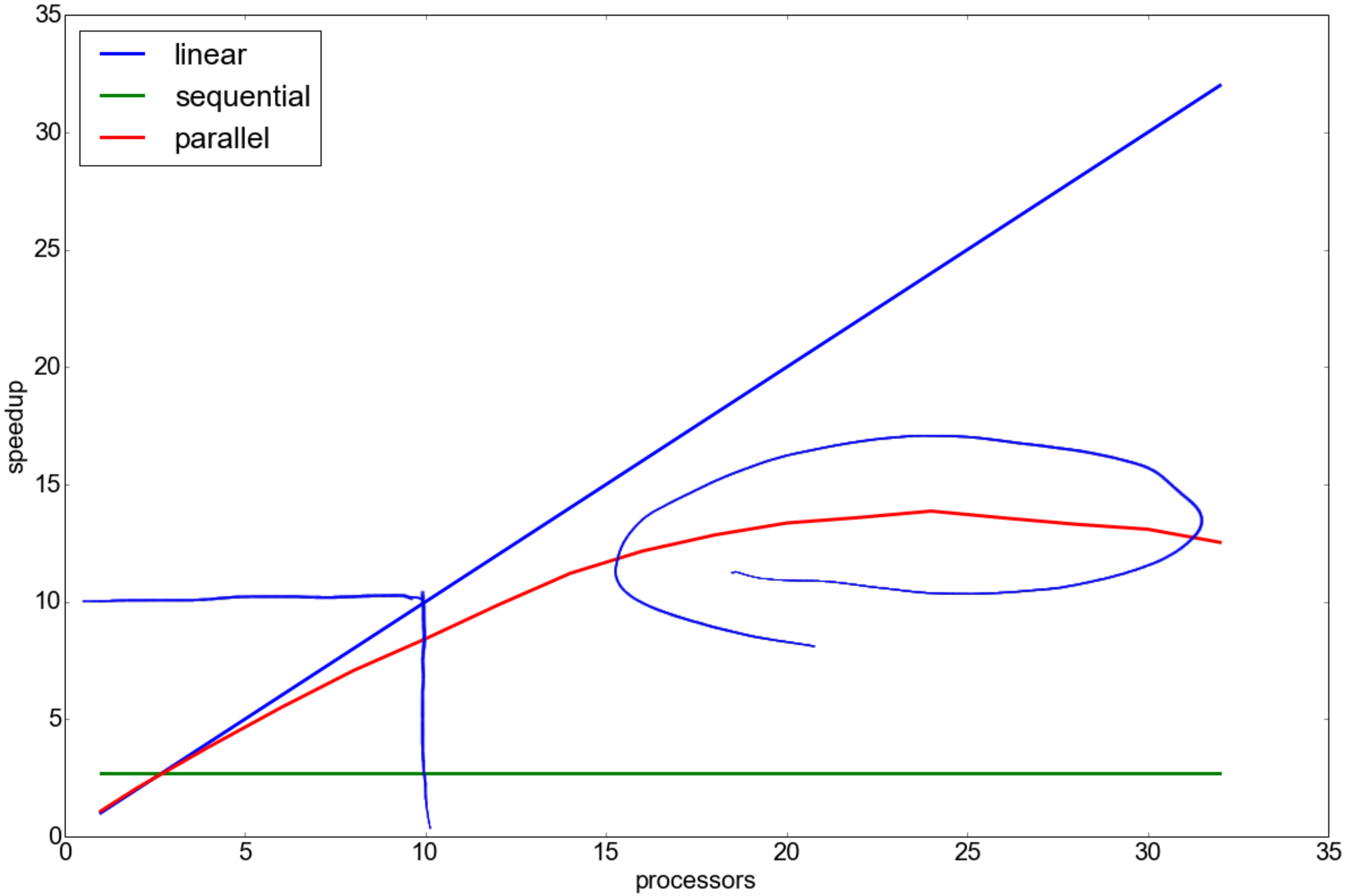
(Baseline: serialization of the parallel algorithm).

Sometimes there is a better serial algorithm that does not parallelize well.

In these cases it is fairer to use that algorithm for T_1 (absolute speedup).

Using an unnecessarily poor baseline artificially inflates speedup and efficiency.

(parallel) speedup graph example



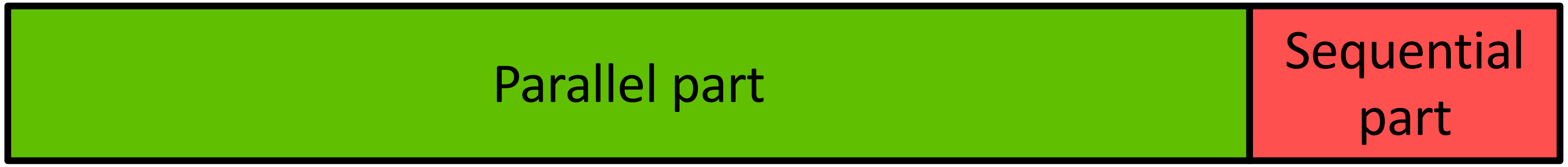
why $S_p < p$?

- Programs may not contain enough parallelism
  - e.g., some parts of program might be sequential

- overheads introduced by parallelization
  - typically associated with synchronization

- architectural limitations
  - e.g., memory contention

Question:



Parallel program:

- sequential part: 20%
- parallel part: 80% (assume it scales linearly)
- $T_1 = 10$

What is T_8 ? What is the speedup S_8 ?

$T_1 = 10$

seq : $10 \cdot 0.2$

par : $10 \cdot 0.8$

$p = 8$

$T_8 = ?$

$$T_8 = T_{seq} + T_{par}$$

$$= 10 \cdot 0.2 + \frac{10 \cdot 0.8}{8}$$

$$= 2 + 1$$

$$T_8 = 3$$

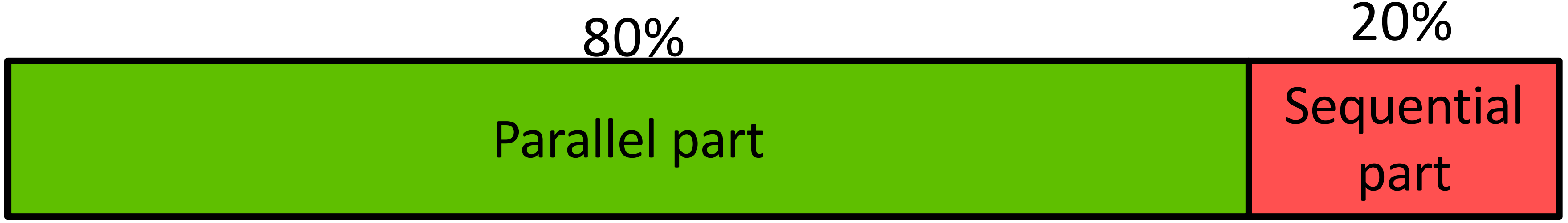
$$S_8 = \frac{T_1}{T_8} = \frac{10}{3} = 3.\overline{3}$$

$$E = \frac{S_8}{p} = \frac{3.\overline{3}}{8} = 0.4 \quad 40\%$$

80% // 20% seq

P	T_1	T_p	$S_p = \frac{T_1}{T_p}$	$E = \frac{S_p}{p}$
1	10	10	1	1
2	10	$2 + \frac{8}{2} = 6$	$\frac{10}{6} = 1.\overline{6}$	$\frac{1.\overline{6}}{2} = 83\%$
3	10	$2 + \frac{8}{3} = 4.\overline{6}$	$\frac{10}{4.\overline{6}} \approx 2.1$	$\frac{2.1}{3} \approx 71\%$

Answer:



- $T_1 = 10$
- $T_8 = 3$
- $S_8 = T_1/T_8 = 10/3 = 3.33$

Amdahl’s Law

…the effort expended on achieving high parallel processing rates is wasted unless it is accompanied by achievements in sequential processing rates of very nearly the same magnitude.

— Gene Amdahl

Amdahl's Law – Ingredients

Execution time T_1 of a program falls into two categories:

- Time spent doing non-parallelizable serial work
- Time spent doing parallelizable work

Call these W_{ser} and W_{par} respectively

Amdahl's Law – Ingredients

Given P workers available to do parallelizable work, the times for sequential execution and parallel execution are:

$$T_1 = \underline{W_{ser}} + \underline{W_{par}}$$

And this gives a bound on speed-up:

$$T_p \geq \underline{W_{ser}} + \boxed{\frac{W_{par}}{P}}$$

Amdahl's Law

Plugging these relations into the definition of speedup yields Amdahl's Law:

$$S_p \leq \frac{W_{ser} + W_{par}}{W_{ser} + \frac{W_{par}}{P}}$$

$S_p = \frac{T_1}{T_p}$

Amdahl's Law - Corollary

$$S_p \le \frac{W_{ser} + W_{par}}{W_{ser} + \frac{W_{par}}{P}}$$

If $\boldsymbol{f}$ is the non-parallelizable serial fractions of the total work, then the following equalities hold:

$$W_{ser} = \boldsymbol{f}T_1,$$
$$W_{par} = (1 - \boldsymbol{f})T_1$$

which gives:

$$S_p \le \frac{1}{\boldsymbol{f} + \frac{1 - \boldsymbol{f}}{P}}$$

$$S_p \le \frac{w_{ser} + w_{par}}{w_{ser} + \frac{w_{par}}{p}} = \frac{f\cancel{T_1} + (1-f)\cancel{T_1}}{f\cancel{T_1} + \frac{(1-f)\cancel{T_1}}{p}}$$

$$= \frac{1}{f + \underbrace{\frac{(1-f)}{p}}_{\to 0}} \qquad p \to \infty$$

AMDAHL

$p \to \infty$

$$S_\infty \le \frac{1}{f}$$

Bsp: $f = 0.1$ ($\to$ 10% sequ.)

$$S_\infty \le \frac{1}{0.1} = 10 \times$$

$f = 0.01$ ($\to$ 1% sequ.)

$$S_\infty \le \frac{1}{0.01} = 100 \times$$

GUSTAFSON

$$S_p \le \frac{f\cancel{T} + p(1-f)\cancel{T}}{f\cancel{T} + \frac{\cancel{p}(1-f)\cancel{T}}{\cancel{p}}}$$

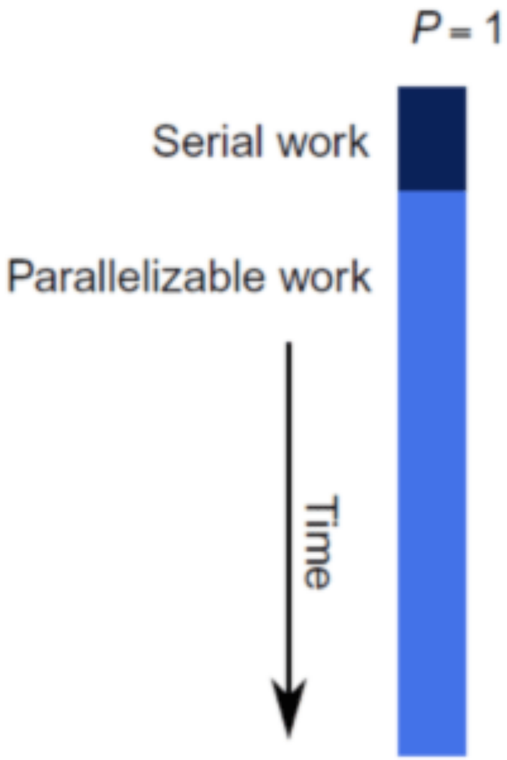
$$\le \frac{f + p(1-f)}{\underbrace{f + (1-f)}_{1}}$$

$$S_p \le f + p(1-f)$$

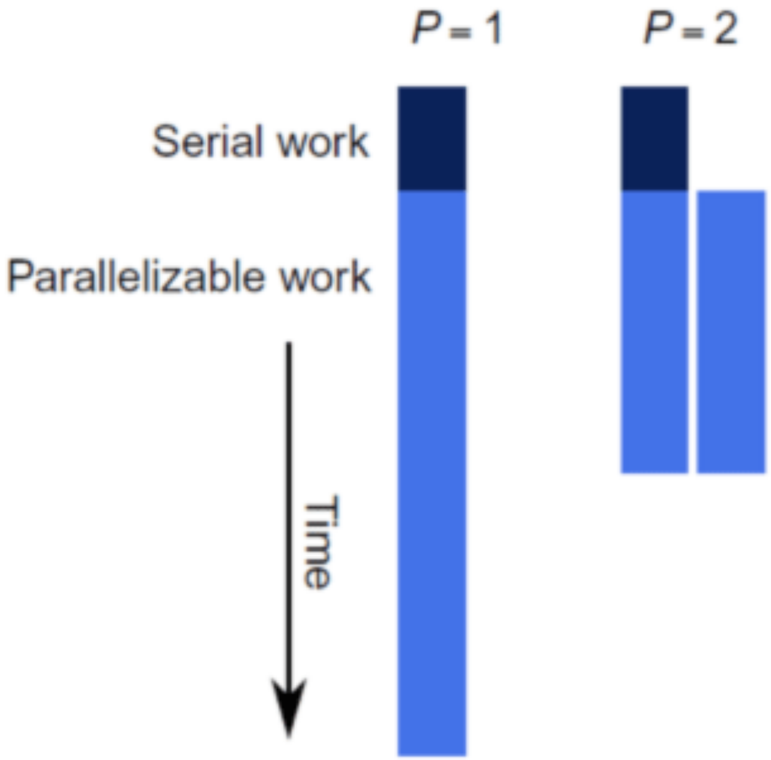
What happens if we have infinite workers?

$$S_\infty \leq \frac{1}{f}$$

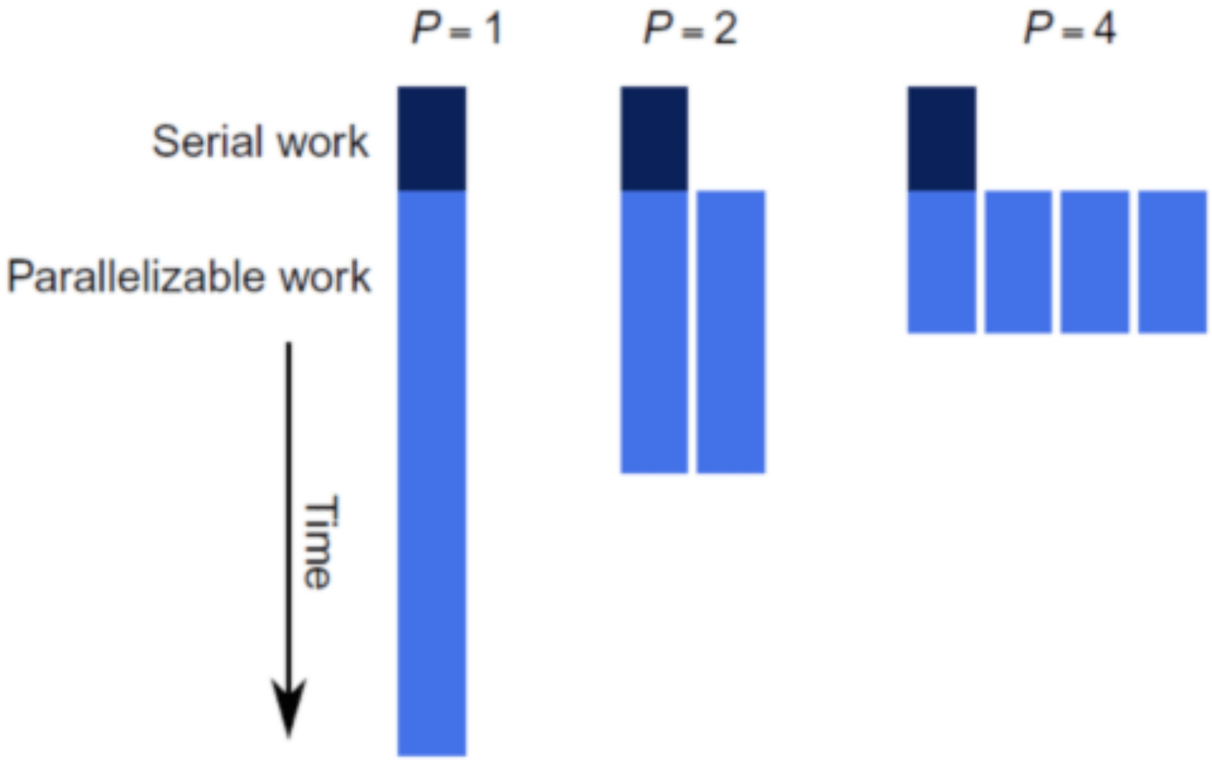
Amdahl's Law Illustrated



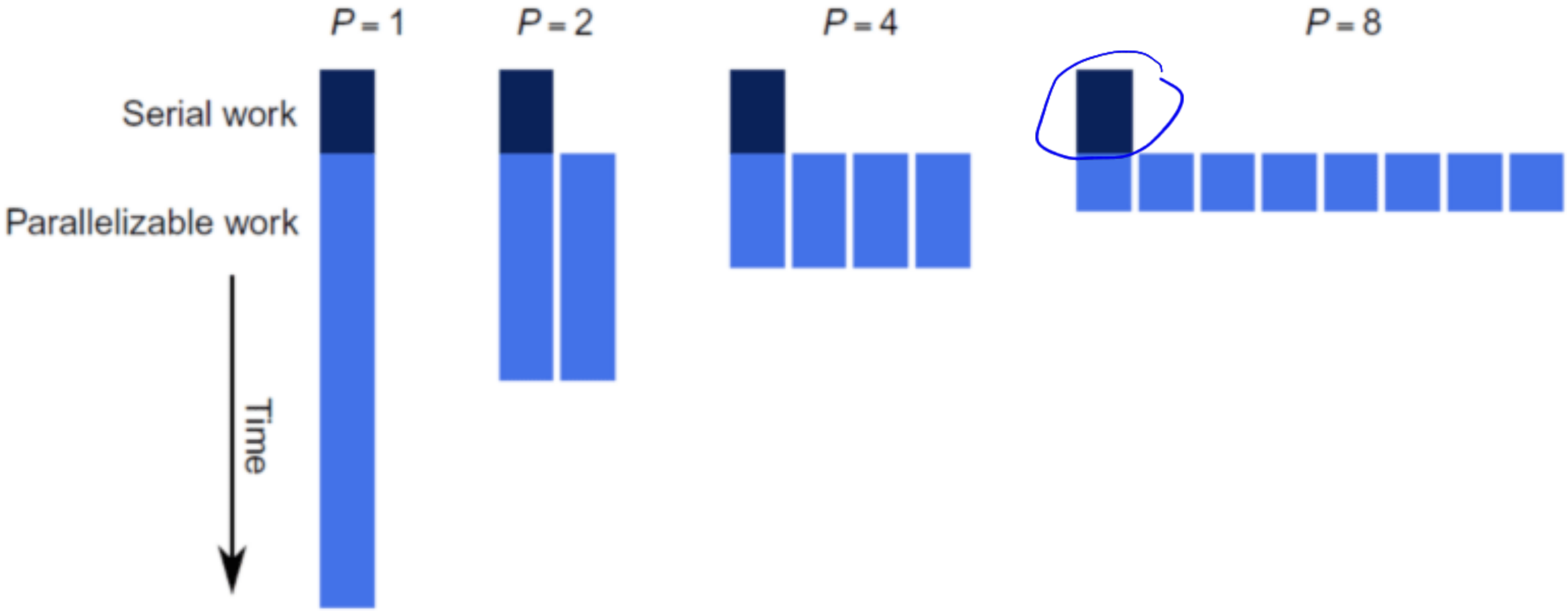
Amdahl’s Law Illustrated



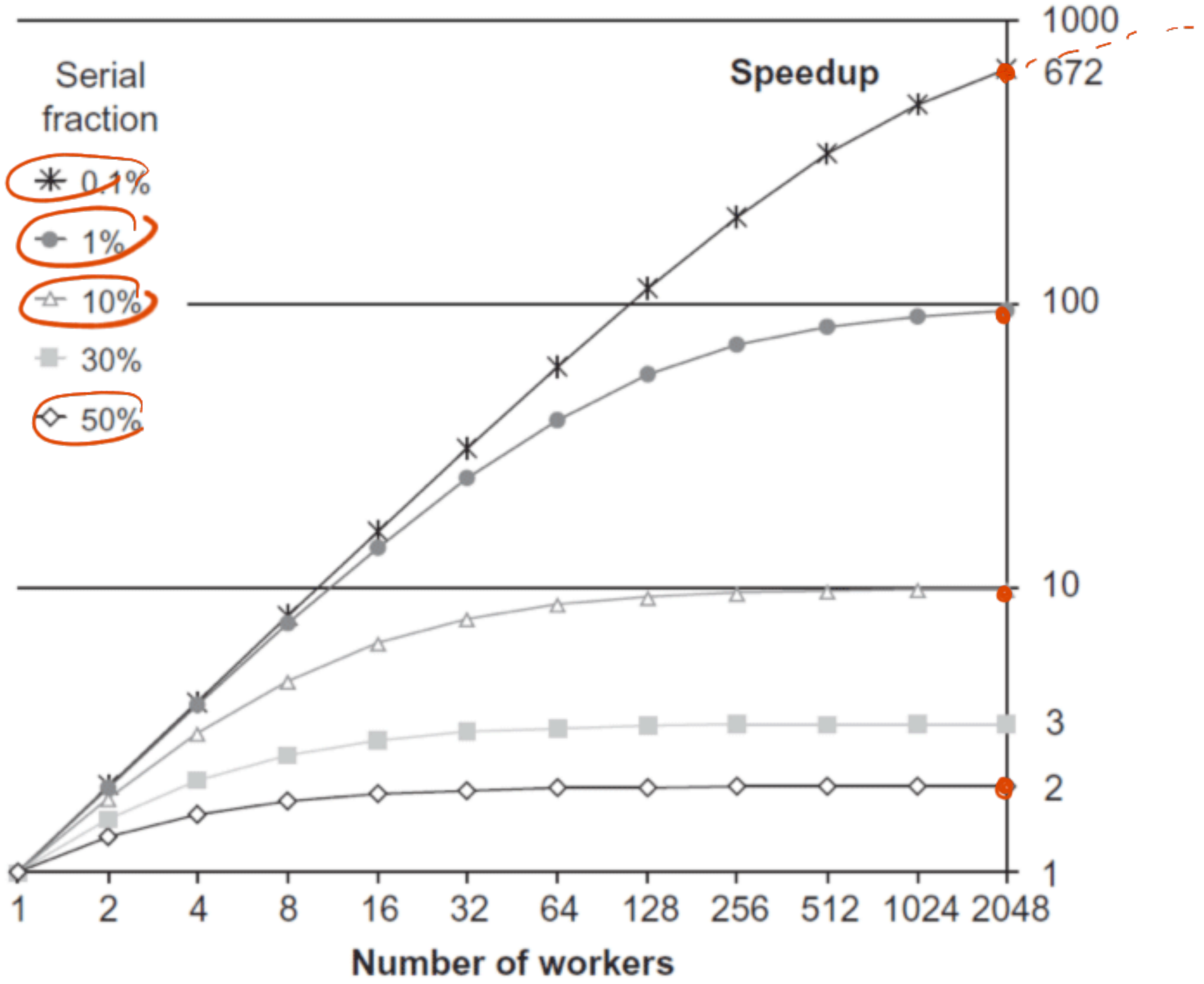
Amdahl’s Law Illustrated



Amdahl's Law Illustrated

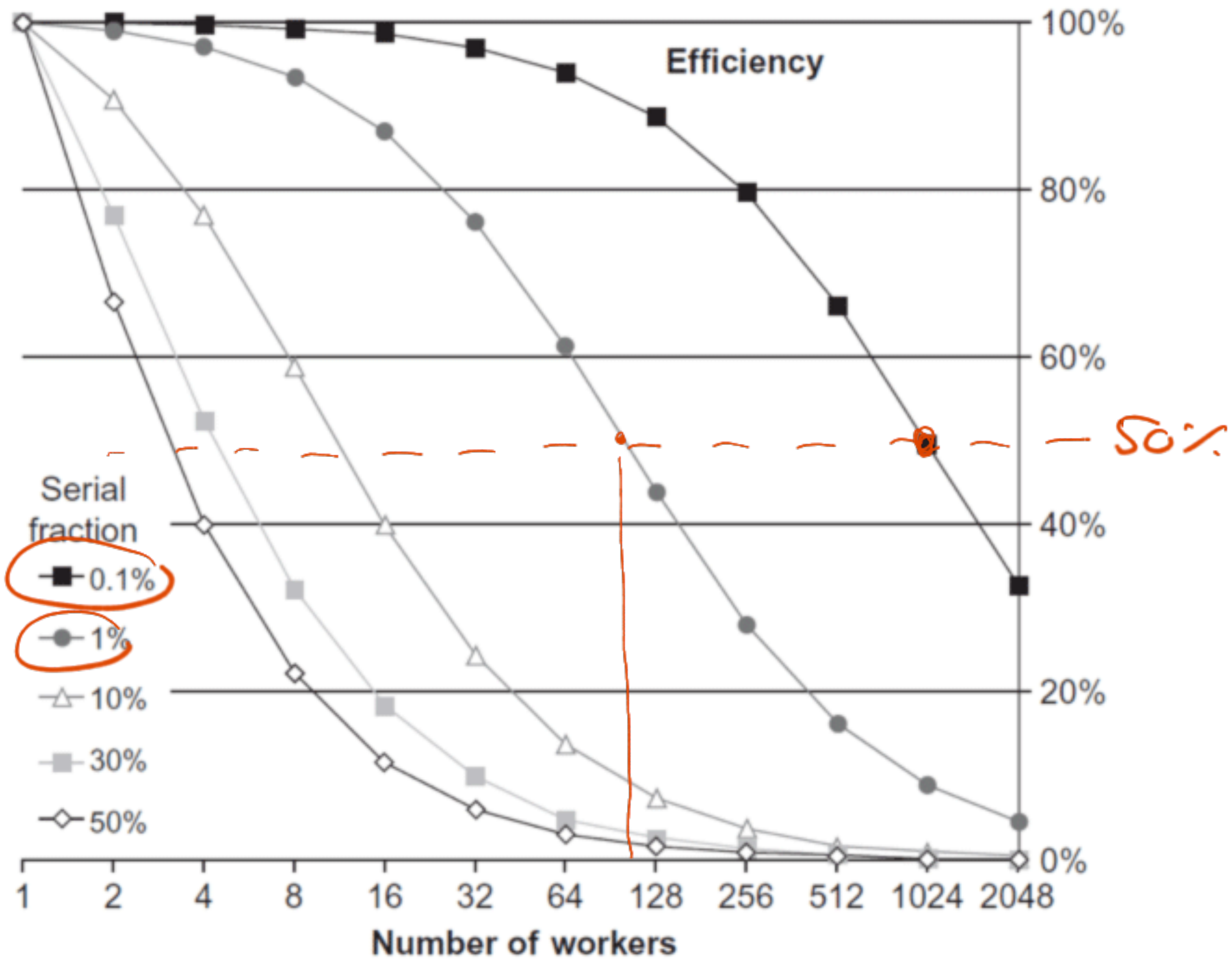


Speedup



Serial fraction
0.1%
1%
10%
30%
50%
Speedup
1000
672
100
10
3
2
1
1
2
4
8
16
32
64
128
256
512
1024
2048
Number of workers

Efficiency



Efficiency
100%
80%
60%
40%
20%
0%
50%
Serial fraction
0.1%
1%
10%
30%
50%
1
2
4
8
16
32
64
128
256
512
1024
2048
Number of workers

Remarks about Amdahl's Law

- It concerns ***maximum speedup*** (Amdahl was an optimist *(or pessimist?)*)
  - architectural constraints will make factors worse
- But his law is ***mostly bad news*** (as it puts a limit on scalability)

- takeaway: **all non-parallel parts of a program (no matter how small) can cause problems**

- Amdahl's law shows that efforts required to further reduce the fraction of the code that is sequential may pay off in large performance gains.

- Hardware that achieves even a small decrease in the percent of things executed sequentially may be considerably more efficient

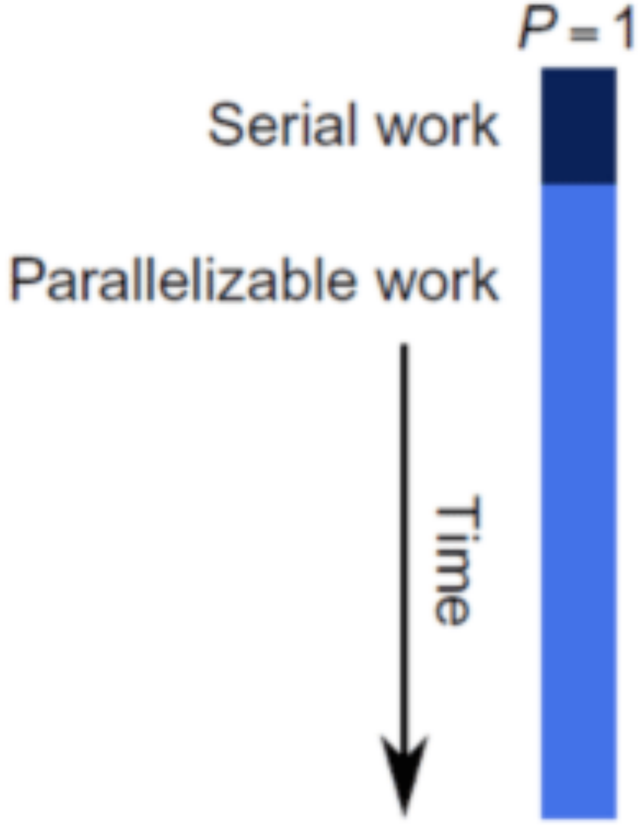
Gustafson's Law

- An alternative (optimistic) view to Amdahl's Law

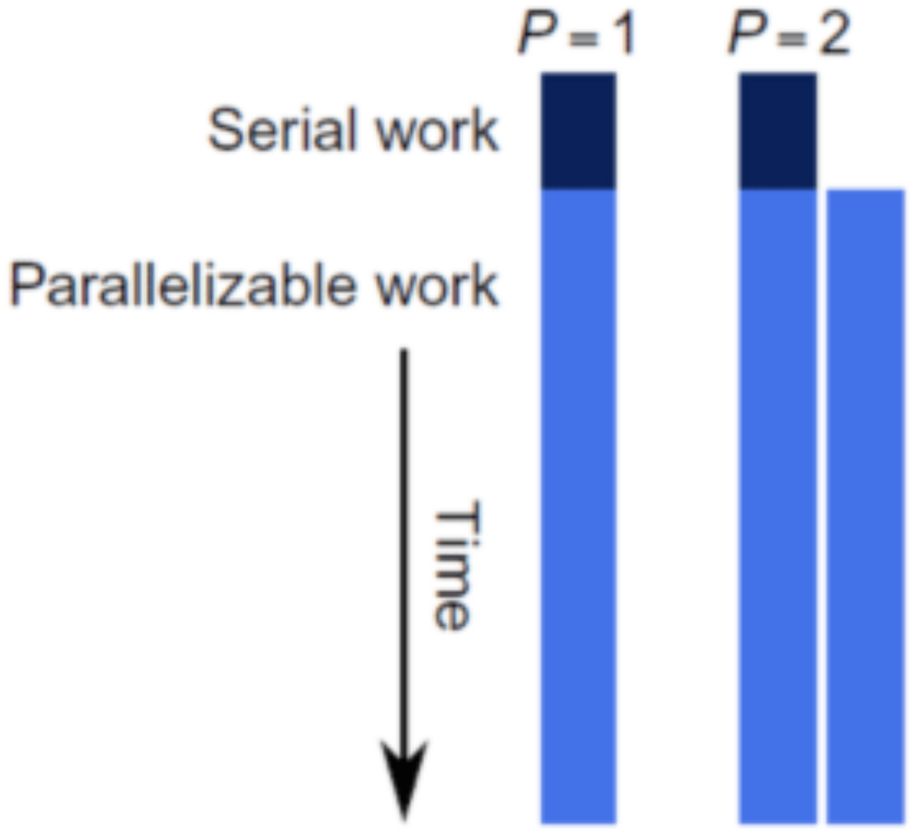
Observations:

- consider problem size
- run-time, not problem size, is constant
- more processors allows to solve larger problems in the same time
- parallel part of a program scales with the problem size

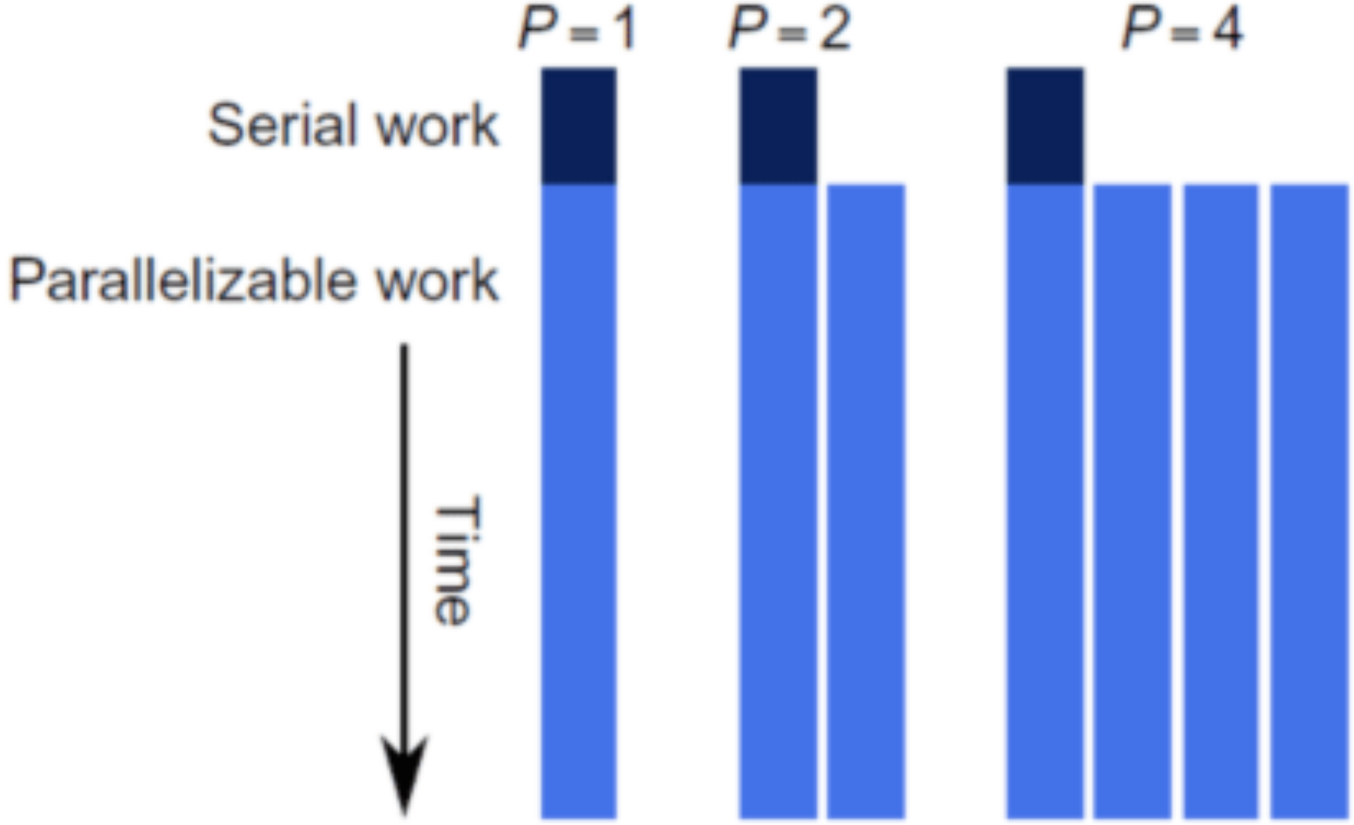
Gustafson’s Law



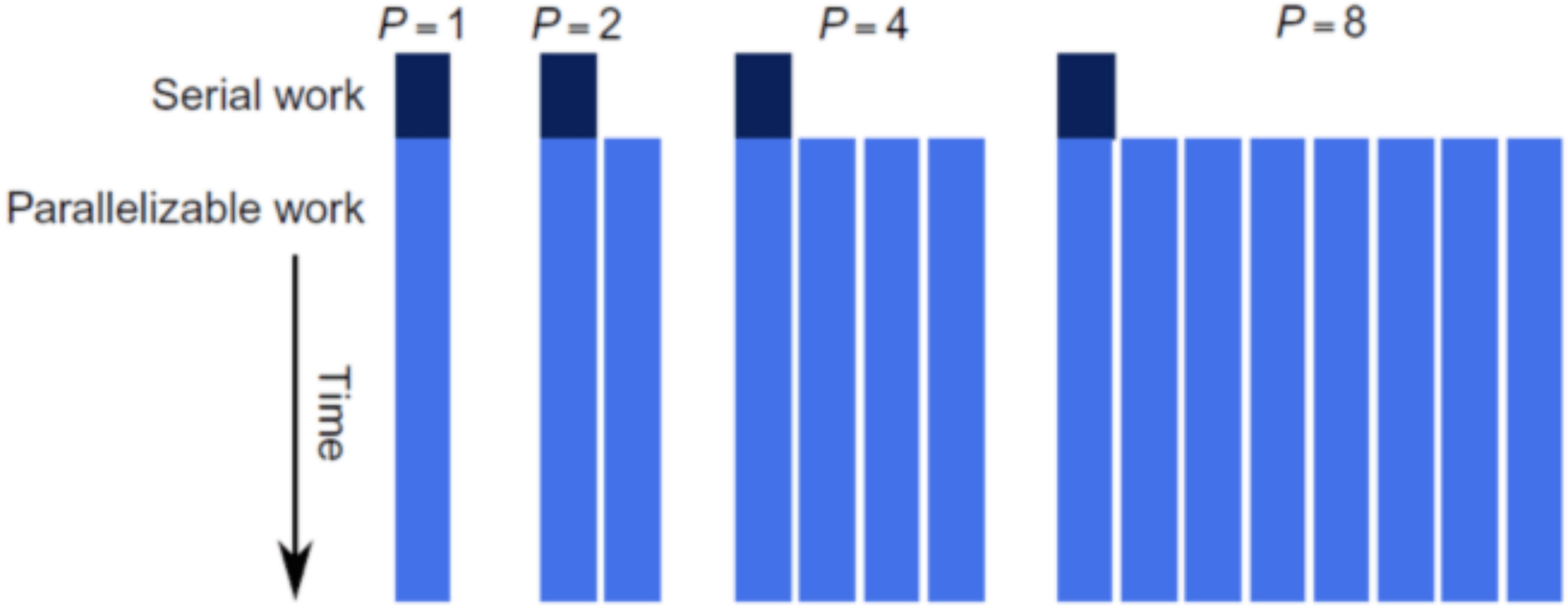
Gustafson’s Law



Gustafson’s Law



Gustafson's Law



Gustafson's Law

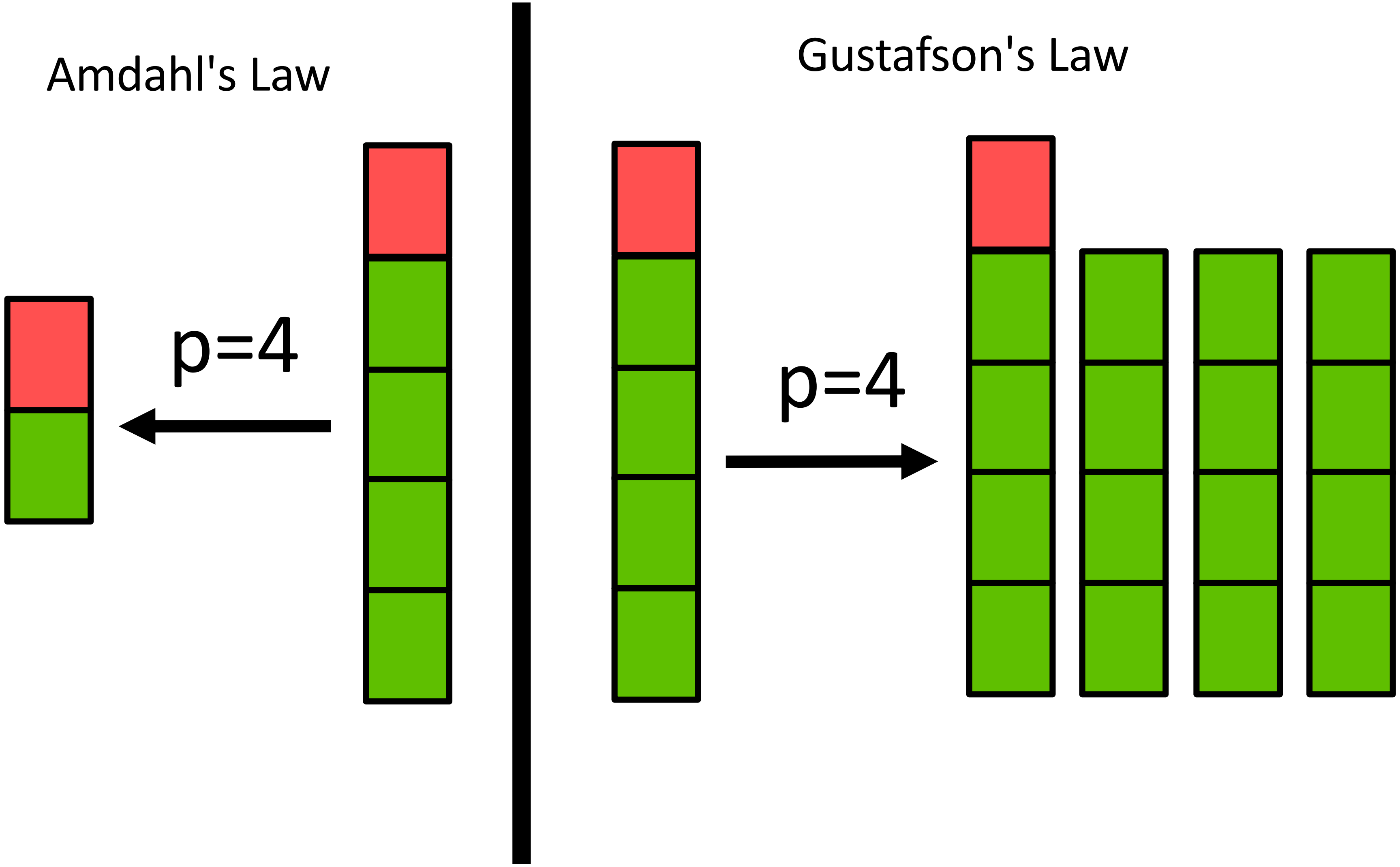
- f: sequential part (no speedup)

$$W = p(1-f)T_{wall} + fT_{wall}$$

$$S_p = f + p(1-f)$$
$$= p - f(p-1)$$

http://link.springer.com/referenceworkentry/10.1007%2F978-0-387-09766-4_78

Amdahl's vs Gustafson's Law



Summary

- Parallel speedup
- Amdahl's and Gustafson's law
- Parallelism: task/thread granularity