



Memory Systems

Counteracting the memory wall

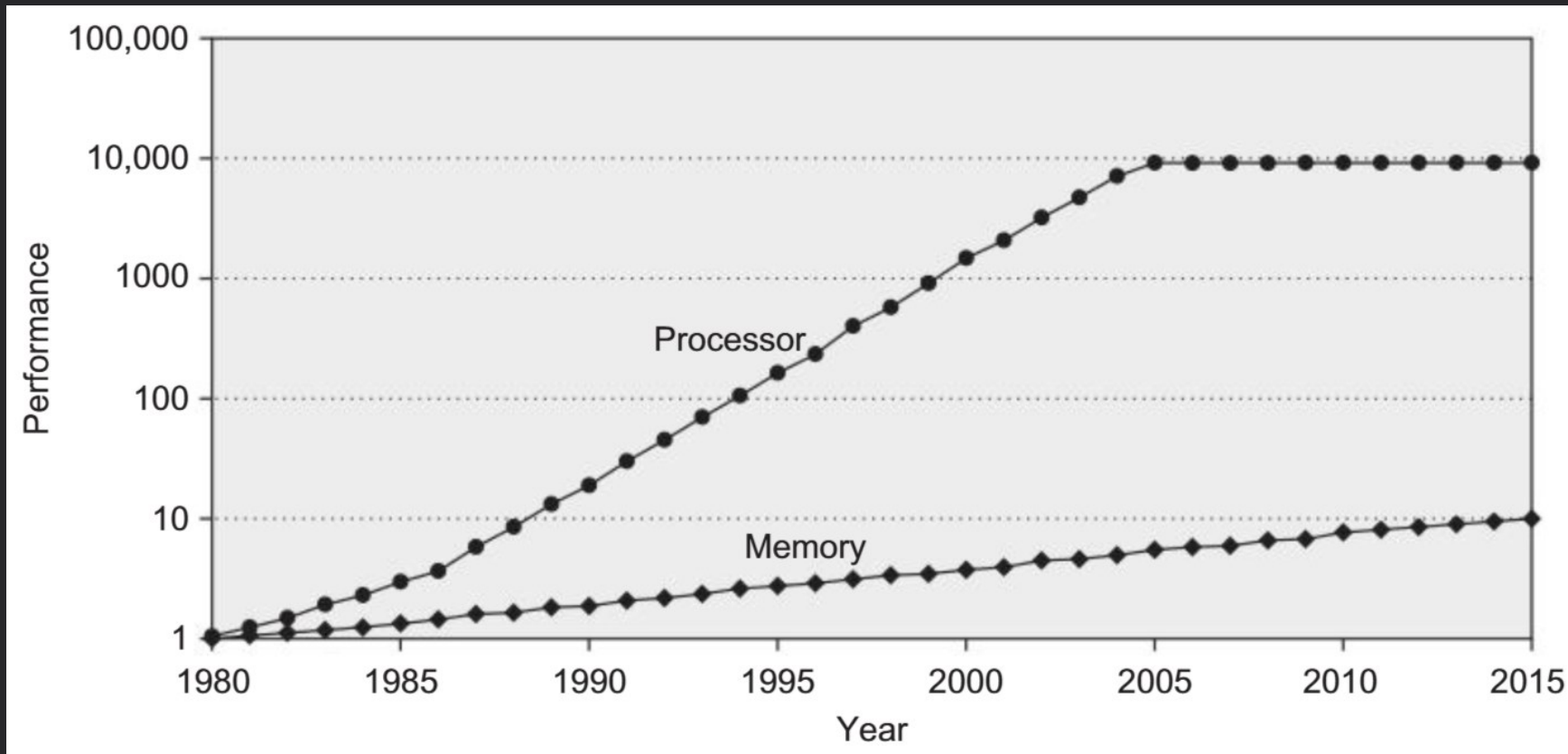
Introduction

Analogy

Evaluation

Conclusion

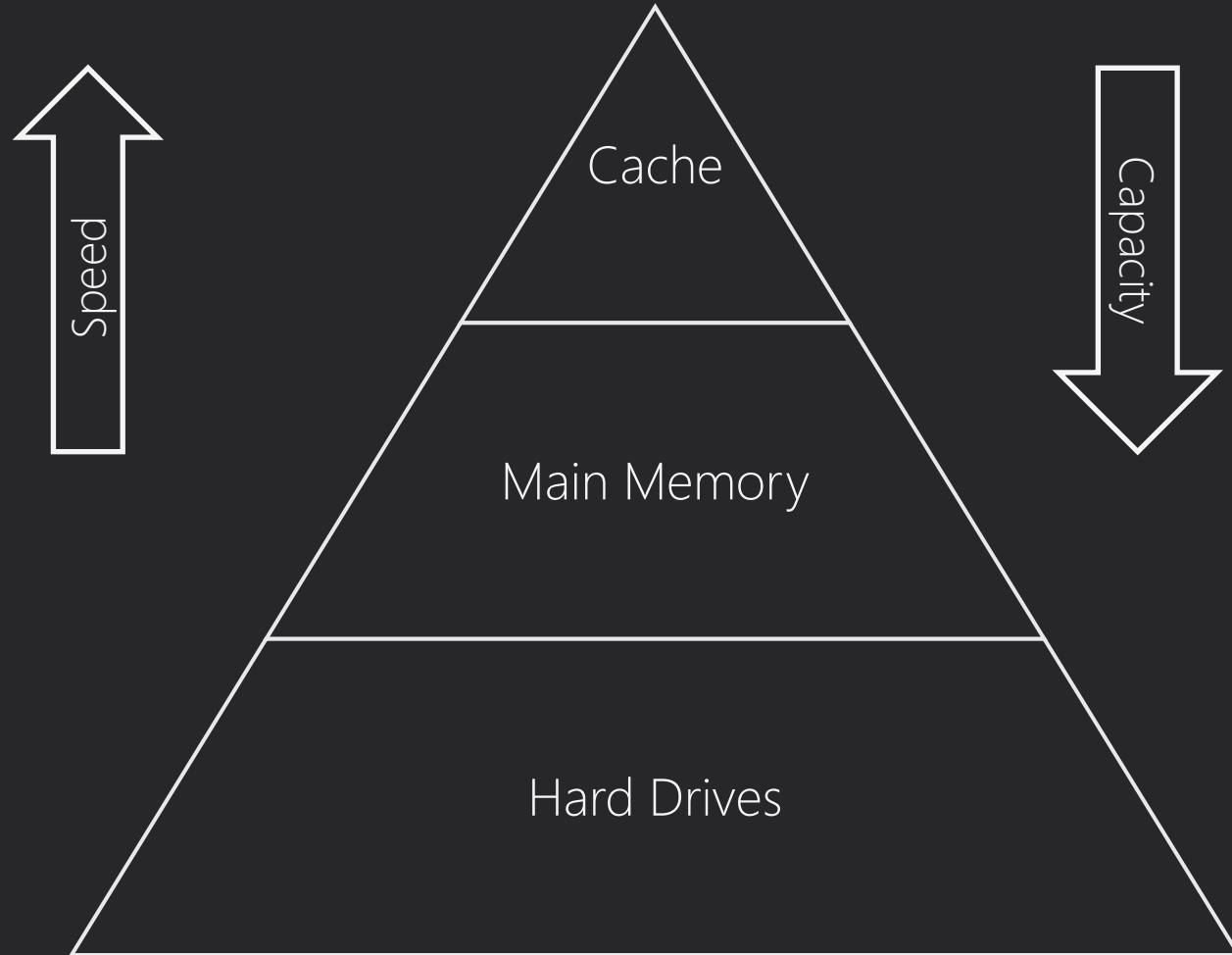
The memory wall



John L. Hennessy and David A. Patterson. "Computer Architecture: A Quantitative Approach, 6th ed." Morgan Kaufmann, 2017

- DRAM performance improvement
 - 7% per year
- Processor performance improvement
 - 30 to 50% per year until 2005

The memory hierarchy



	Technology	Cost (\$) / GB	Location
Cache	SRAM	Very expensive	On chip
Memory	DRAM	Inexpensive	Off chip
Hard drive	Flash	Cheap	Off chip

| An analogy: the library



I Temporal and spatial locality

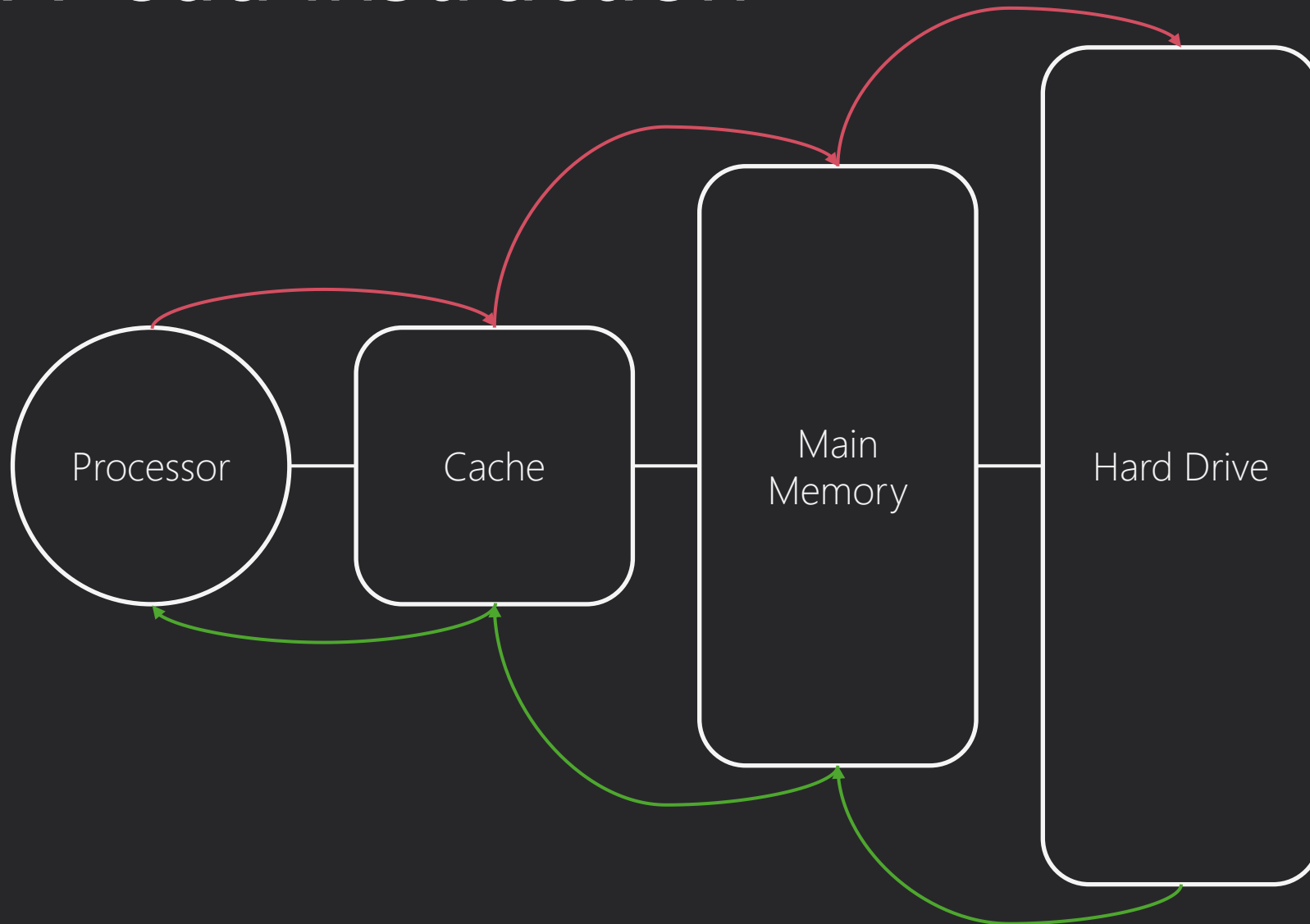
- Temporal locality
 - If you referenced a book recently, you will probably reference it again soon
- Spatial locality
 - If you pulled one book off the shelf, other books on that shelf are also relevant to you

Memory system evaluation



Performance metrics and equations

A load instruction



- Each memory component has some latency to access
 - $t_{cache}, t_{MM}, t_{HD}$
- Classifying accesses
 - A *miss*: the data requested is not available
 - A *hit*: the data requested is available

The hit and miss rate

$$\text{Miss Rate} = \frac{\text{Number of misses}}{\text{Total number of memory accesses}}$$


$$\text{Hit Rate} = \frac{\text{Number of hits}}{\text{Total number of memory accesses}}$$

Let MR_{cache} , MR_{MM} be the miss rates of the cache and main memory, respectively.

■ Average memory access time

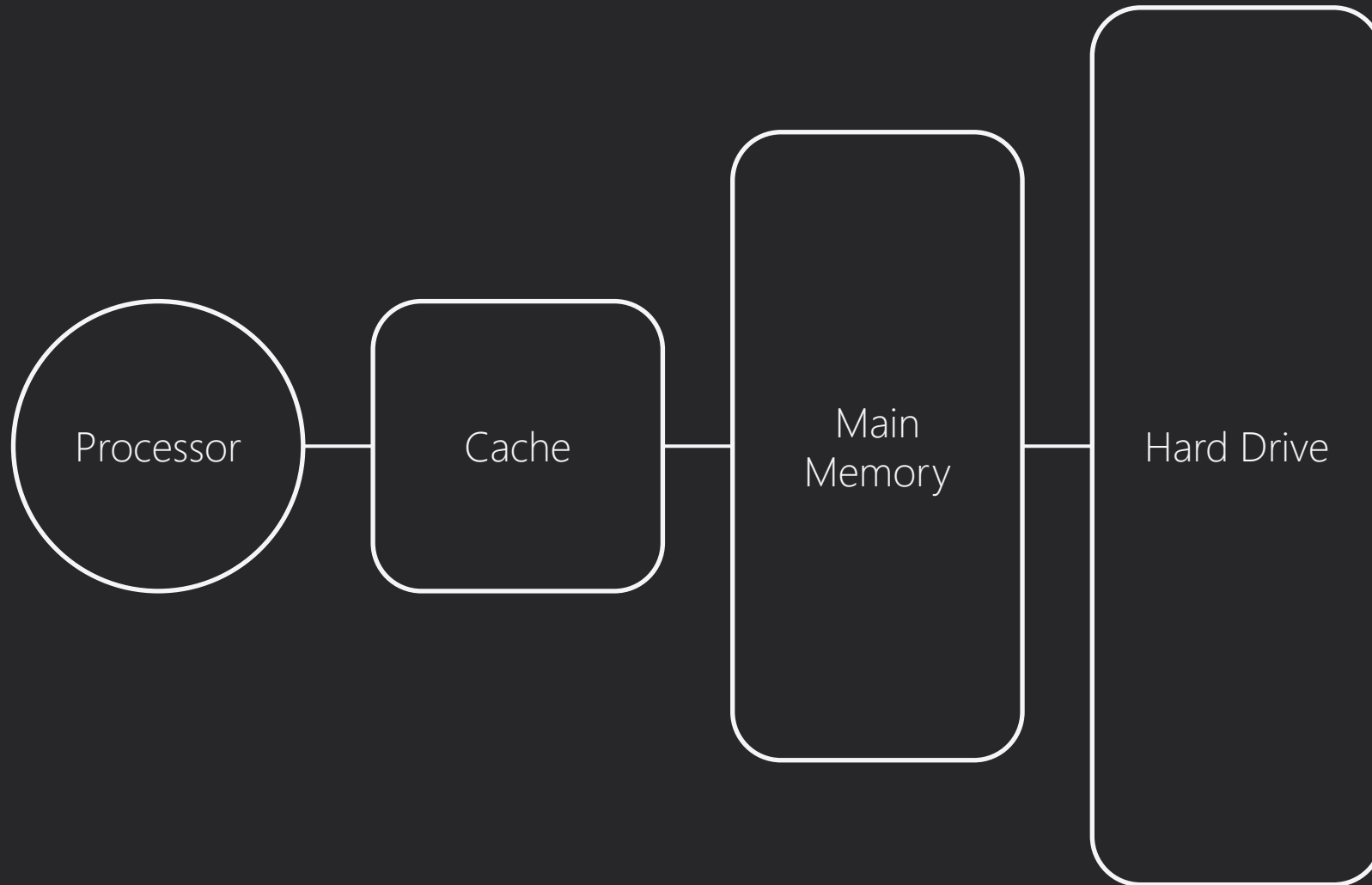
$$AMAT = t_{cache} + MR_{cache}(t_{MM} + MR_{MM} \times t_{HD})$$

Conclusion



Recapping the important points

| The memory hierarchy



The cache hierarchy

