Lecture III - Caching

Programming: Everyday Decision-Making Algorithms

Dr. Nils Roemer
Kühne Logistics University Hamburg - Winter 2024

Learning Objectives

By the end of this lecture, you will be able to:

- Explain the fundamental concepts of caching and its importance
- Compare different cache replacement strategies
- Identify caching principles in everyday life
- Apply caching concepts to personal productivity
- Understand the relationship between caching and attention management

Introduction

Caching: Making the most of limited resources

> In the practical use of our intellect, forgetting is as important a function as remembering. – [William James]

- Today's lecture is on caching.
- The discussed topics are highly relevant for everything that has to do with computers and data.
- On top of that, caching is another very important concept for our everyday decision-making.

Let's approach the topic using an everyday decision-making situation

- We have a problem: Our cupboard.
- It's time to put things in order.

Question: What could we do?

- Better organization
- Clearing out things we no longer need
- Now we have two problems:
  - Storing?
  - Clearing out?

Storing

How to bring order to storage?

- Subdivide storage
- Efficient sorting
- Increase capacity

Subdivide storage and efficient sorting outcome

- Time investment can improve storing.
- That is a difficult trade-off in itself.
- Nontheless: Each storage has a limited capacity.

Question: What do we do, when the storage is full?

We could increase the capacity

But…

- Increase capacity is costly.
- There is a trade-off between size and speed.
- Sooner or later, every storage will fill up if not cleared out.

Question: What types of storages can you imagine that are affected by this?

- Our cupboard
- Our computer (hard drive, RAM, cache, …)
- Our brain?!

Question: What is the impact of a full storage?

- Access speed drops significantly
- Processing time increases
- Overall performance decreases

Clearing out

Now we see, why clearing out is so important

- And that goes for our cupboard as well as for computers and other storages…
- But what stays and what goes?

Question: What replacement/eviction policies can you imagine?

- Random Eviction
- First-In, First-Out (FIFO)
- Least Recently Used (LRU)

Once again, we can learn a lot from the computer sciences.

The evolution of computer memory

- In the 1950s, computer science faced the same question…
- …and has faced it repeatedly since then.
- Processors have become faster and faster (Moore's Law).
- The demands on memory also grew.

CPU

- The processor (CPU, Central Processing Unit) is central to a computer and is often referred to as the "brain" of the system.
- It executes instructions and performs calculations, to process data and run programs.

Problem: Access Time

Problem: No matter how much faster the processor gets, if input data isn't available fast enough or can't be stored quickly enough, the system won't become faster overall.

Question: What is the solution?

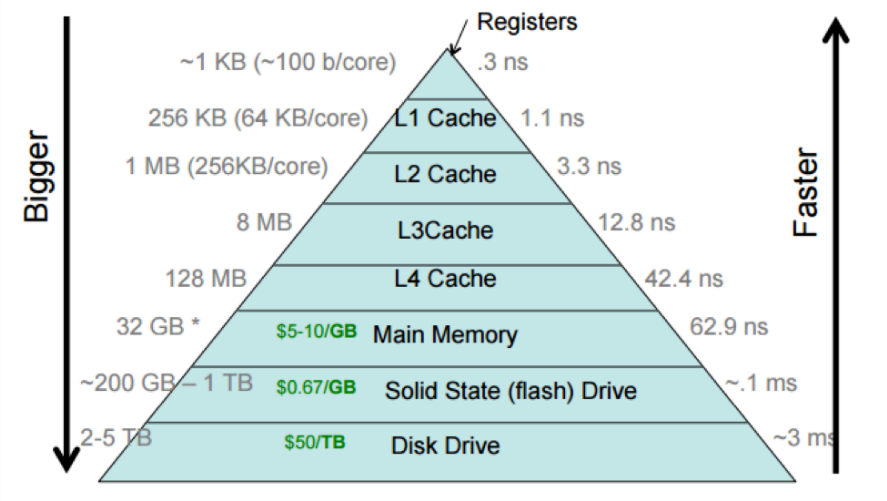
Cache

Cache

Hierarchical memory pyramid for efficient data management

- L1 Cache, directly on CPU, very fast access time.
- L2 Cache, between L1 and main memory.
- RAM memory.
- Mass Storage (hard drive).
- The whole memory system works like a library.
- Find an explanation here.

Cache - Size and Speed trade-off



Registers are 10 million times faster than the hard drive!

Library Principle

- Library storage (5 million books, Mass Storage)
- Subject locations (100K books, RAM)
- Desk (5 borrowed books, L2)
- Short-term memory (L1)

Library Principle

- L1 and L2 cache only contain most necessary data.
- The same should apply to your desk.
- Therefore, both must be cleared regularly.

Clearing out Strategies

How to clear up?

- Optimal: Clairvoyance
  - Store everything in the cache that will be needed
  - Delete everything from the cache that won't be needed
- Question: What is the problem with this strategy?
- Optimal strategy not achievable in reality

Realisable strategies?

- Least recently used is the dominant strategy.
- Evicts the least recently accessed item from the cache when space is needed.
- Leads to much better performance on average than, for example, random eviction.
- Question: Why do you think least recently used is the better strategy?

Managerial and personal insights:

- Let go of things you haven't used in ages
- Keep things where they are used
- Both have been proven to contribute to a significant increase in productivity

Keeping things where they are used...

Might be optimal, in a mathematical sense

Productivity

The strong limitations of caches make them a “security risk”.

- Denial-of-Service Attacks (DoS) attacks.
  - Cache Flooding
  - Cache Poisoning
- Overload a system with excessive requests or data.
- Causing it to slow down or crash.
- The system is forced to evict important data.

Question: Why are the findings about cache so important for humans?

Your Brain is a Cache, Not a Database -[Joe Chrysler]

Our brain has similar weaknesses

Productivity and creativity are negatively affected by:

- Overload (too much)
- Exhaustion (too long)
- Context switching (interruption of “flow”, 23 minutes to get back on track)
- Distraction (Cache Flooding)
- Fake News (Cache Poisoning)

This can lead to burnout.

Question: What do you think should we do about it?

Stolen Focus: Why You Can’t Pay Attention - Key Insights I

Main Idea:

- Modern life is eroding our ability to concentrate and engage deeply, with societal factors affecting personal focus.
- The reason for that has much to do with the way our brain works and how we interact with technology.

Stolen Focus: Why You Can’t Pay Attention - Key Insights II

Key Causes of Attention Loss:

- Technology & Social Media: Designed to capture and keep attention, leading to fragmented focus.
- Constant Interruptions: Notifications and multitasking disrupt deep work and concentration.
- Environmental & Lifestyle Factors: Poor sleep, stress, and diet impact cognitive function.

Stolen Focus: Why You Can’t Pay Attention - Key Insights III

Consequences:

- Reduced ability to think critically and creatively.
- Difficulty sustaining attention on meaningful tasks.

Stolen Focus: Why You Can't Pay Attention - Key Insights IV

Solutions Suggested:

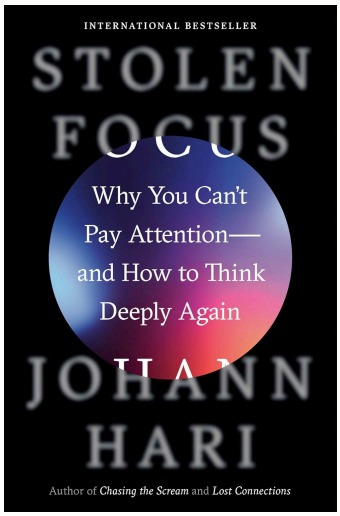
- Limit screen time and practice "monotasking."
- Prioritize sleep, nutrition, and mindful habits.

Stolen Focus: Why You Can't Pay Attention - Key Insights V

https://www.youtube.com/watch?v=DqlywBxYELw

But even better: read the book!

Stolen Focus: Why You Can't Pay Attention - Book



Mitigation

Mitigation

- Distraction can hardly be avoided in today's world but can be mitigated.
- This is particularly important for managers.
- This lecture is designed to raise your awareness of what you can do to keep your brain working efficiently.

Awareness I

Train awareness

- Spotlight – immediate goals – Focus
- Starlight – medium-term goals – Wishes
- Daylight – long-term goals – Values

Awareness II

Prioritization

- Prioritization
- Structure (Schedule)
- Breaks

- Enable flow (dedicated workspace, manage notifications, clear communication)
- Meditation & exercise

Key Takeaways

Key Takeaways

- Caching is a universal concept that applies to computers, organizations, and human cognition
- Efficient cache management requires strategic decisions about what to keep and what to remove
- LRU (Least Recently Used) is often the most practical replacement strategy
- Our brain's limitations are similar to computer caches
- Managing our attention and focus requires understanding these limitations

Summary Quiz

Question: Take a moment to reflect:

1. What are the three main types of cache replacement strategies?
2. How does the library principle relate to computer memory hierarchy?
3. What are two ways you can apply caching principles to improve your productivity?
4. How can understanding cache flooding help protect against information overload?

Literature

Interesting literature to start

- Christian, B., & Griffiths, T. (2016). Algorithms to live by: the computer science of human decisions. First international edition. New York, Henry Holt and Company.[1]
- Ferguson, T.S. (1989) 'Who solved the secretary problem?', Statistical Science, 4(3). doi:10.1214/ss/1177012493.

Books on Programming

- Downey, A. B. (2024). Think Python: How to think like a computer scientist (Third edition). O'Reilly. Here
- Elter, S. (2021). Schrödinger programmiert Python: Das etwas andere Fachbuch (1. Auflage). Rheinwerk Verlag.

...

> **i** Note
>
> Think Python is a great book to start with. It's available online for free. Schrödinger Programmiert Python is a great alternative for German students, as it is a very playful introduction to programming with lots of examples.

[1] The main inspiration for this lecture. Nils and I have read it and discussed it in depth, always wanting to translate it into a course.

More Literature

For more interesting literature, take a look at the literature list of this course.