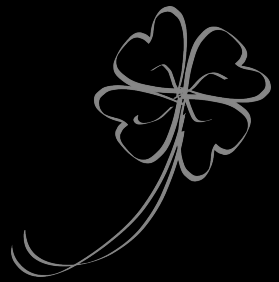
Fifty Shades of Caching

and how LLMs Paint it Black

effie mouzeli
[[User:Effie_Mouzeli_(WMF)]]



SREcon25 EMEA
Dublin 2025
@manjiki.bsky

WIKIMEDIA
FOUNDATION

Fifty Shades of Caching

and how LLMs Paint it Bl

effie mouzeli
[[User:Effie_Mouzeli_(WMF)]]

SREcon25 EMEA
Dublin 2025
fosstodon.org/@manjiki

WIKIMEDIA
FOUNDATION

Fifty Shades of Caching

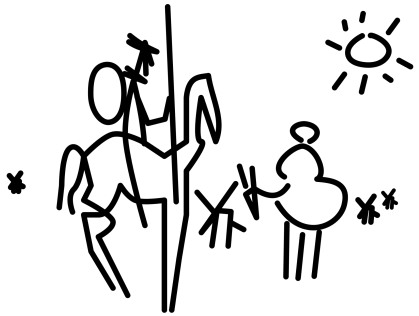
and how LLMs Paint it Bl

effie mouzeli
[[User:Effie_Mouzeli_(WMF)]]

SREcon25 EMEA
Dublin 2025

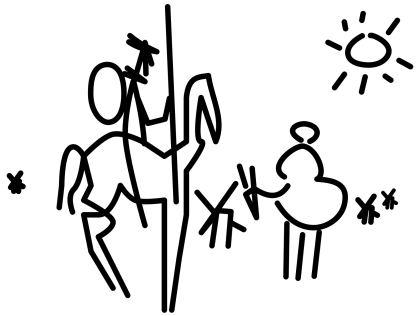
meta.wikimedia.org/wiki/User:Effie_Mouzeli_(WMF)

About



DISCLAIMER

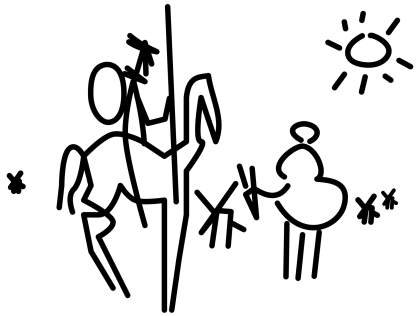
WIKIMEDIA
FOUNDATION



DISCLAIMER

No LLMs were harmed in the making of this talk

WIKIMEDIA FOUNDATION



DISCLAIMER

Views are my own and not my employer’s

WIKIMEDIA FOUNDATION


WIKIMEDIA
TOPLULUĞU KULLANICI GRUBU
TÜRKİYE
tr.wikimedia.org
WIKIMEDIA
HACKATHON
WIKIMEDIA

Did you know...

- ... the infrastructure hosting Wikipedia (and sister projects) is run by the **Wikimedia Foundation,** an American nonprofit charitable organisation?
- ... all content is managed by volunteers?
- ... we support 300+ languages?
- ... Wikipedia is 25 years old?
- ... Wikipedia has been training LLMs since they were just called ‘statistics.’
- ... we host some truly extraordinary content

WIKIMEDIA
FOUNDATION

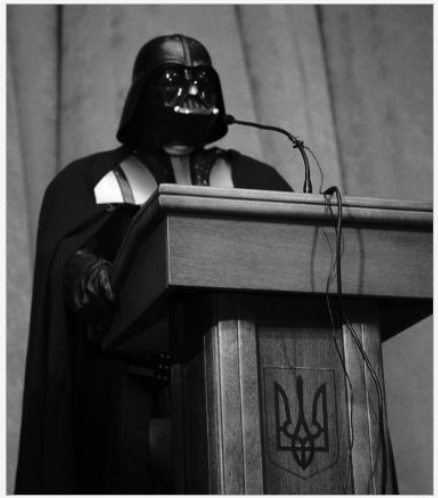
Darth Vader in Ukrainian politics

3 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Since 2012, individuals adopting the name of the *Star Wars* character Darth Vader have applied for national and local government positions in Ukraine.[1] Candidates named Darth Vader (Ukrainian: Дарт Вейдер) tried running for offices in several local, presidential, and parliamentary elections. "Vader" also participated in political and social activism, notably in the Odesa region.



"Darth Alekseevich Vader", a candidate in 2014 Ukrainian presidential election

History [edit]

On 14 February 2011, a man dressed as Darth Vader applied to the Odesa City Council for 10 acres of land for free. According to him, he learned that the council was allocating land on the coastal slopes for free, and came to get a plot for himself, claiming that "the city council members, the executive

Lord Buckethead

8 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Not to be confused with Buckethead.

Lord Buckethead is a novelty candidate who has stood in four British general elections since 1987, portrayed by several individuals. He poses as an intergalactic villain resembling the *Star Wars* character Darth Vader.

Lord Buckethead was created by the American filmmaker Todd Durham for his 1984 science fiction film *Hyperspace*. Without authorisation, Mike Lee stood as Lord Buckethead in the 1987 UK general election and again in the 1992 general election. The character went unused until the comedian Jonathan Harvey stood as Lord Buckethead in the 2017 general election. His televised appearance standing next to prime minister Theresa May went viral, drawing media coverage and an online following.

Following the 2017 election, Durham asserted his ownership of Lord Buckethead and displaced Harvey. With Durham's authorisation, Lord Buckethead returned in 2019, now played by David Hughes. He appeared at People's Vote rallies calling for a second Brexit referendum, and stood in the 2019 general election representing the Monster Raving Loony Party. Harvey continues to campaign using his own character, Count Binface.

Lord Buckethead

Lord Buckethead in 2020

First appearance *Hyperspace* (1984)

Glossary of physics

8 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

This **glossary of physics** is a list of definitions of terms and concepts relevant to physics, its sub-disciplines, and related fields, including mechanics, materials science, nuclear physics, particle physics, and thermodynamics. For more inclusive glossaries concerning related fields of science and technology, see Glossary of chemistry terms, Glossary of astronomy, Glossary of areas of mathematics, and Glossary of engineering.

Contents:

A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S · T · U · V · W · X · Y · Z · See also · References · External links

A [edit]

ab initio

A mathematical model which seeks to describe atomic nuclei by solving the non-relativistic Schrödinger equation for all constituent nucleons and the forces that exist between them. Such methods yield precise results for very light nuclei but become more approximate for heavier nuclei.

Abbe number

*Also called the **V-number** or **constringence**.*

In optics and lens design, a measure of a transparent material's dispersion (a variation of refractive index versus wavelength). High values of V indicate low dispersion.

absolute electrode potential

In electrochemistry, the electrode potential of a metal measured with respect to a universal reference system (without any additional metal–solution interface).

Part of a series on
Physics

Index · Outline · Glossary
History (timeline)
Branches [show]
Research [show]
Physics portal · Category
V • T • E

Wikipedia:You *do* need to cite that the sky is blue

7 languages

Project page | Talk | Read | Edit | View history | Tools

From Wikipedia, the free encyclopedia

This is an essay.
It contains the advice or opinions of one or more Wikipedia contributors. This page is not an encyclopedia article, nor is it one of Wikipedia's policies or guidelines, as it has not been thoroughly vetted by the community. Some essays represent widespread norms; others only represent minority viewpoints.

Shortcuts
WP:NOTBLUE
WP:DOCITEBLUE
WP:NOTBLUESKY
WP:REDSKY
WP:SKYISNOTBLUE

This page in a nutshell: Just because something appears obvious to you, doesn't mean it's obvious to everyone. Build articles from reliable, expert sources, and cite those sources.

It is sometimes felt that "obvious" statements, such as "the sky is blue", do not need citing. However, there are some reasons why you do need to cite the "obvious", such as that the sky is blue.

First of all, you do need citations in the "main" article, i.e., where the subject is the "obvious" statement or its major element. I.e., the statement "the sky is blue" must be footnoted in the article "Sky", especially in the section that discusses the color of the sky. Such references usually lead to more detailed knowledge.

The color of the sky will vary depending on time of day, local

Wikipedia:You don't need to cite that the sky is blue

8 languages

Project page | Talk | Read | Edit | View history | Tools

From Wikipedia, the free encyclopedia

"WP:BLUE", "WP:BLUESKY", and "WP:SKYBLUE" redirect here. For the blue link color, see H:LC. For the guideline about two or more adjacent links, see MOS:SEAOFBLUE. For the protection levels, see WP:BLUELOCK and WP:SKYBLUELOCK.

This is an essay on the neutral point of view and the verifiability policies.
It contains the advice or opinions of one or more Wikipedia contributors. This page is not an encyclopedia article, nor is it one of Wikipedia's policies or guidelines, as it has not been thoroughly vetted by the community. Some essays represent widespread norms; others only represent minority viewpoints.

Shortcuts
WP:FACTS
WP:BLUE
WP:BLUESKY
WP:OBV
WP:SKYBLUE

This page in a nutshell: Although citing sources is an important part of editing Wikipedia, there is no need to cite information that is already obvious.

Verifiability is an important and core policy of Wikipedia. Article content should be backed up by reliable sources wherever needed to show that the presentation of material on Wikipedia is consistent with the views that are presented in scholarly discourse or the world at large. Such sources help to improve the encyclopedia.

However, many editors misunderstand the citation policy, seeing it as a tool to enforce, reinforce, or cast doubt upon a particular point of view in a content dispute, rather than as a means to verify Wikipedia's information. This can lead to several forms of mildly disruptive editing which are better avoided. Ideally, common sense would always be applied, but site history shows this is unrealistic. Therefore, this essay gives some practical advice.

Personal life

Musk became a U.S. citizen in 2002.[50] From the early 2000s until late 2020, Musk resided in California, where both Tesla and SpaceX were founded.[400] He then relocated to Cameron County, Texas,[401][402] saying that California had become "complacent" about its economic success.[400][403][404]

Musk plays video games, which he stated has a "'restoring effect' that helps his 'mental calibration'".[419] Some games he plays include *Quake*, *Diablo IV*, *Elden Ring*, and *Polytopia*.[420][421] Musk once claimed to be one of the world's top video game players but has since admitted to "account boosting", or cheating by hiring outside services to achieve top player rankings.[422][423][424] Musk has justified the boosting by claiming that all top accounts do it so he has to as well to remain competitive.[425][424][426] In 2024 and 2025, Musk criticized the video game *Assassin's Creed Shadows* and its creator Ubisoft for "woke" content.[427] Musk posted to X that "DEI kills art" and specified the inclusion of the historical figure Yasuke in the *Assassin's Creed* game as offensive; he also called the game "terrible". Ubisoft responded by saying that Musk's comments were "just feeding hatred" and that they were focused on producing a game not pushing politics.[428][429]

About Caching

What is caching

Cache (computing)

From Wikipedia, the free encyclopedia

"Caching" redirects here. For other uses, see Cache (disambiguation).

In computing, a **cache** (/kæʃ/ *KASH*)[1] is a hardware or software component that stores data so that future requests for that data can be served faster; the data stored in a cache might be the result of an earlier computation or a copy of data stored elsewhere. A **cache hit** occurs when the requested data can be found in a cache, while a **cache miss** occurs when it cannot. Cache hits are served by reading data from the cache, which is faster than recomputing a result or reading from a slower data store; thus, the more requests that can be served from the cache, the faster the system performs.[2]

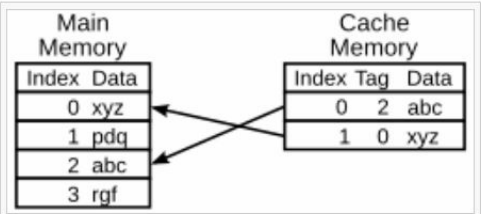


Diagram of a CPU memory cache operation

WIKIMEDIA FOUNDATION

Caching Is Everywhere

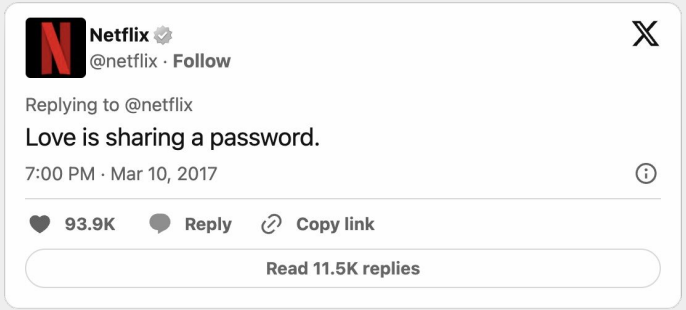
- L1, L2, L3 CPU caches
- Page Caches
- Database query caches
- Distributed Caches
- Content Delivery Networks
- Your ~~Computer~~ Browser

Why is caching important?

Why is caching important?

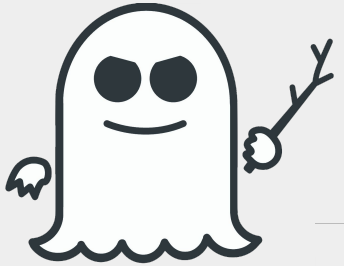
Why is caching important?

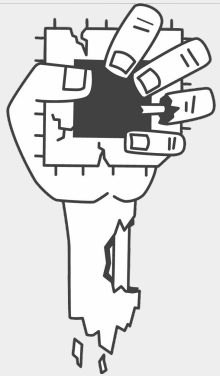
Why is caching important?



Why is caching important?

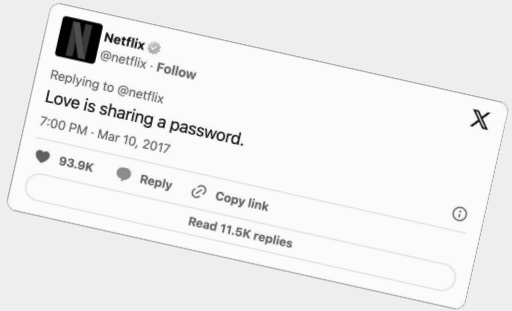
Why is caching important?





Caching systems are optimised to improve performance by analysing human behaviour.

— Plato



About This Talk

Agenda

The CDN Chapter
Caching headers
Browser Cache
CDNs

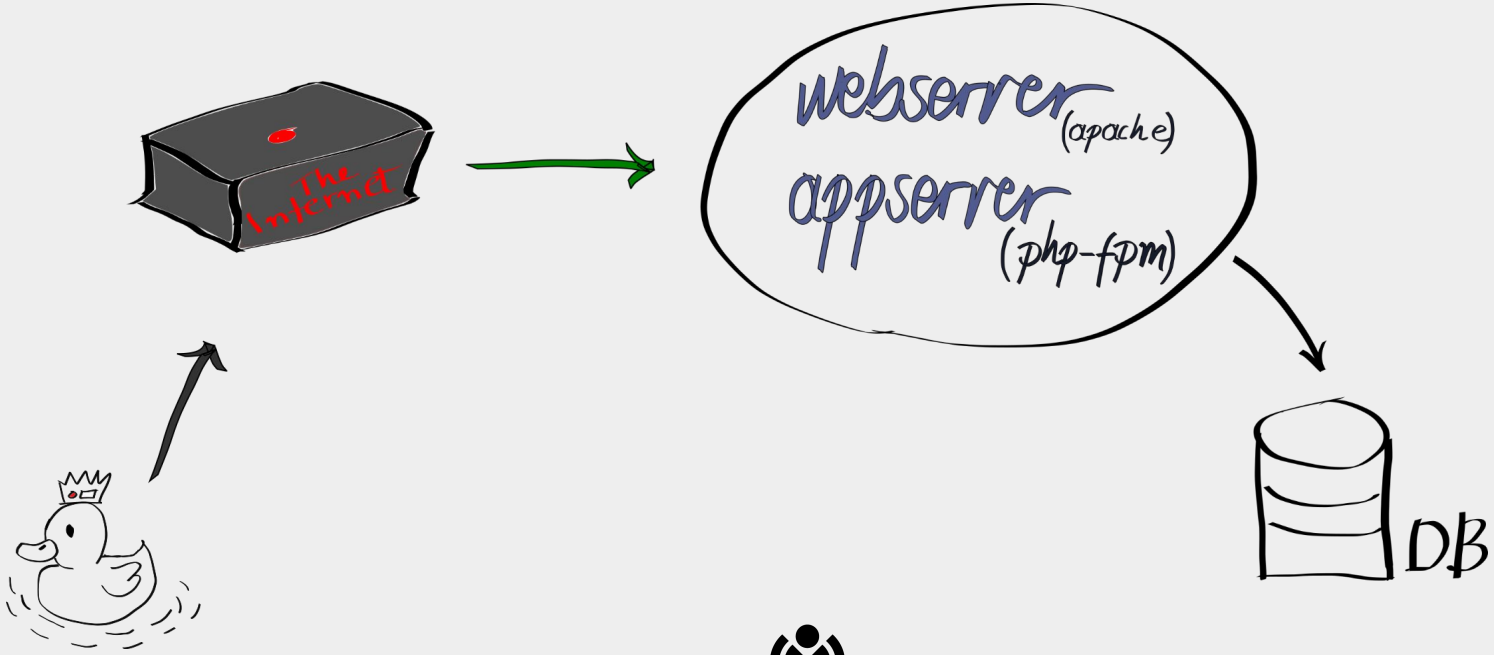
Fifty Shades of Stale
Warm up
Invalidate

Fifty Layers of Application Caching
In-process caches
Distributed caches

Rage Against the Machine Learning
The rise of the machines
Can we strike back?

WIKIMEDIA FOUNDATION

Basic Web Infrastructure



The CDN Chapter

WIKIMEDIA
FOUNDATION

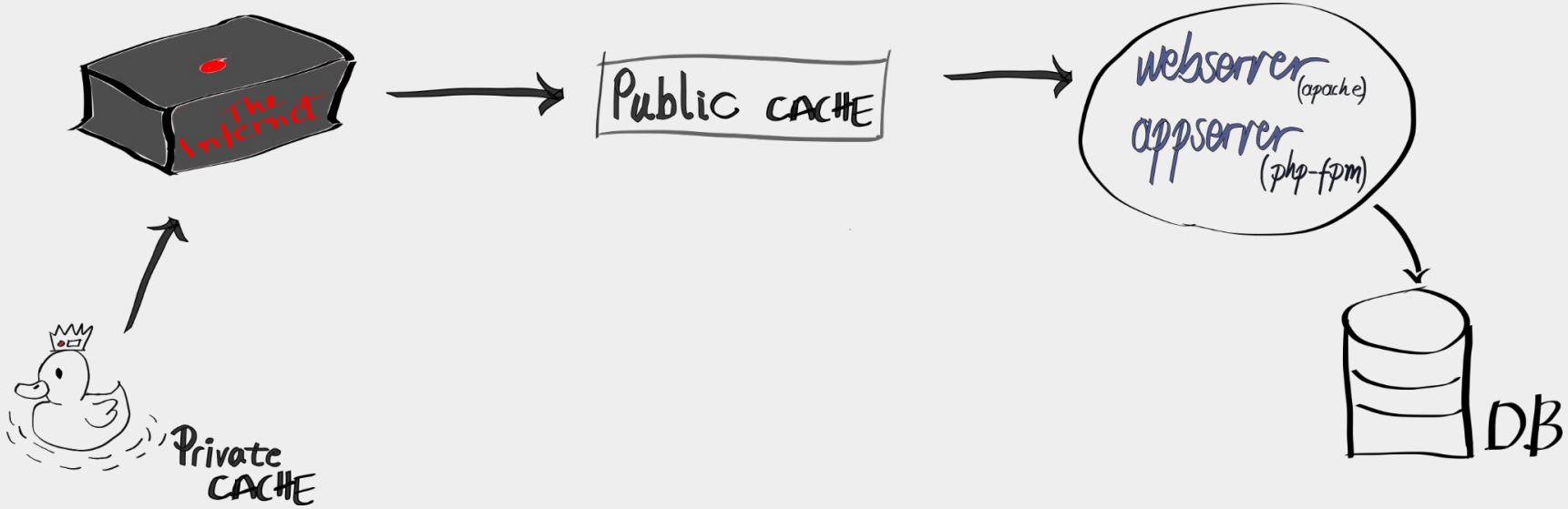
Web Caches

Your Browser ↠ private

Yours or Some Other Proxy ↠ public

Content Delivery Networks ↠ public

Web Caches



WIKIMEDIA FOUNDATION

How to Cache?

- Primary Key: URL
  - URLs with different query params are two different keys
- Varying by HTTP Headers
  - `Accept-Encoding`
  - `Accept-Language`
  - `User-Agent`
- Cache Control Directives
  - `Cache-Control: max-age=3600`
  - `Cache-Control: private`
- Cache Validation
  - `Etag`
  - `Last-Modified`

How to Cache? - Entries

```
Key:GET|https://koko.gr/api/data?id=123|Accept-Language:el-GR|Accept-Encoding:gzip

Response Headers: {...}

Response Body: {...}

Metadata:
 - Cached at: 2025-09-30 14:30:00
 - Expires at: 2025-09-30 15:30:00
 - Max-age: 3600 seconds
 - ETag: "abc123"
```

WIKIMEDIA FOUNDATION

Web Caches

A web cache is just a pile of URL responses with metadata, pretending it's organised.

— Aristotle

How to Cache?

```
$ curl -I https://en.wikipedia.org/static/images/project-logos/enwiki.png

HTTP/2 200
date: Mon, 29 Sep 2025 11:57:40 GMT
etag: "1f0f-63fa4440aaa40"
expires: Tue, 29 Sep 2026 11:57:40 GMT
cache-control: max-age=31536000 ← Good for a year
server: ATS/9.2.11
```

WIKIMEDIA FOUNDATION

How to Cache?

```
$ curl -I https://en.wikipedia.org/w/load.php?lang=en&<...>&skin=vector-2022

HTTP/2 200
date: Tue, 30 Sep 2025 09:11:16 GMT
etag: W/"15hup"
expires: Tue, 30 Sep 2026 09:11:24 GMT
cache-control public, max-age=600, s-maxage=600, stale-while-revalidate=60
server: ATS/9.2.11
Vary: User-Agent
X-Cache: cp3067 hit, cp3067 hit/524
```

WIKIMEDIA FOUNDATION

How to Cache?

```
$ curl -I https://en.wikipedia.org/wiki/Reliability_of_Wikipedia

HTTP/2 200
Date: Mon, 29 Sep 2025 16:00:15 GMT ← When this was cached on the server
Last-Modified: Mon, 29 Sep 2025 14:54:39 GMT ← When this resource changed
Age: 62051 ← ~17hrs
cache-control: private, max-age=0, must-revalidate, no-transform
server: mw-web.codfw.main-7f7f5cb8cb-fx7mb
Vary: Accept-Encoding,Cookie,User-Agent
X-Cache: cp3067 miss, cp3067 hit/42
```

WIKIMEDIA FOUNDATION

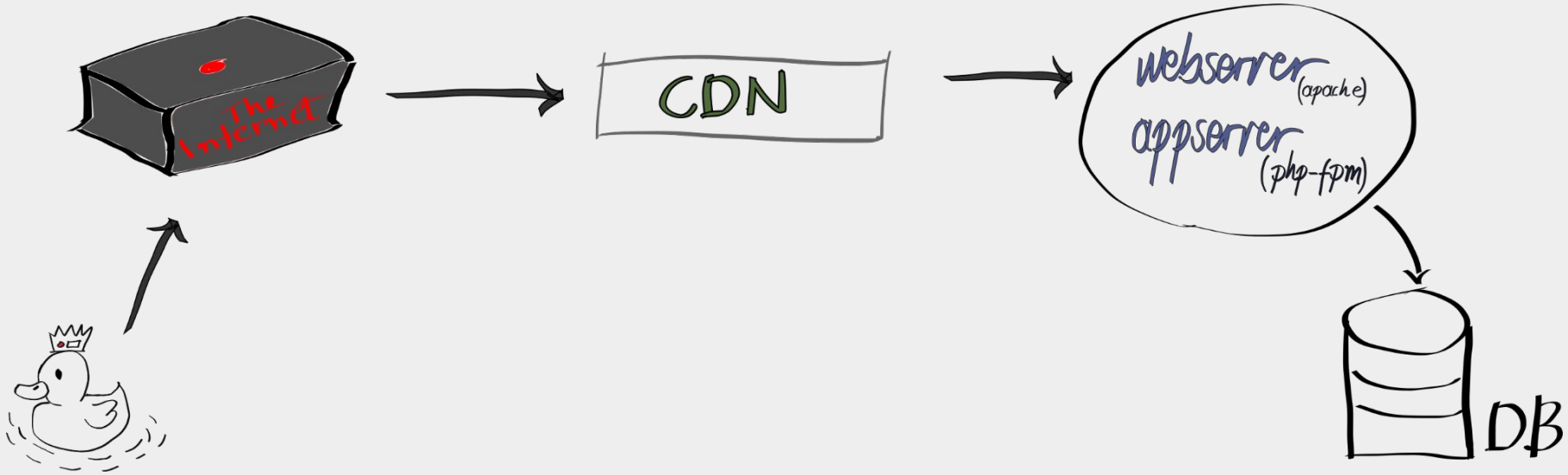
Web Caches

~~Your Browser ↠ private~~

~~Yours or Some Other Proxy ↠ public~~

Content Delivery Networks ↠ public

CDNs

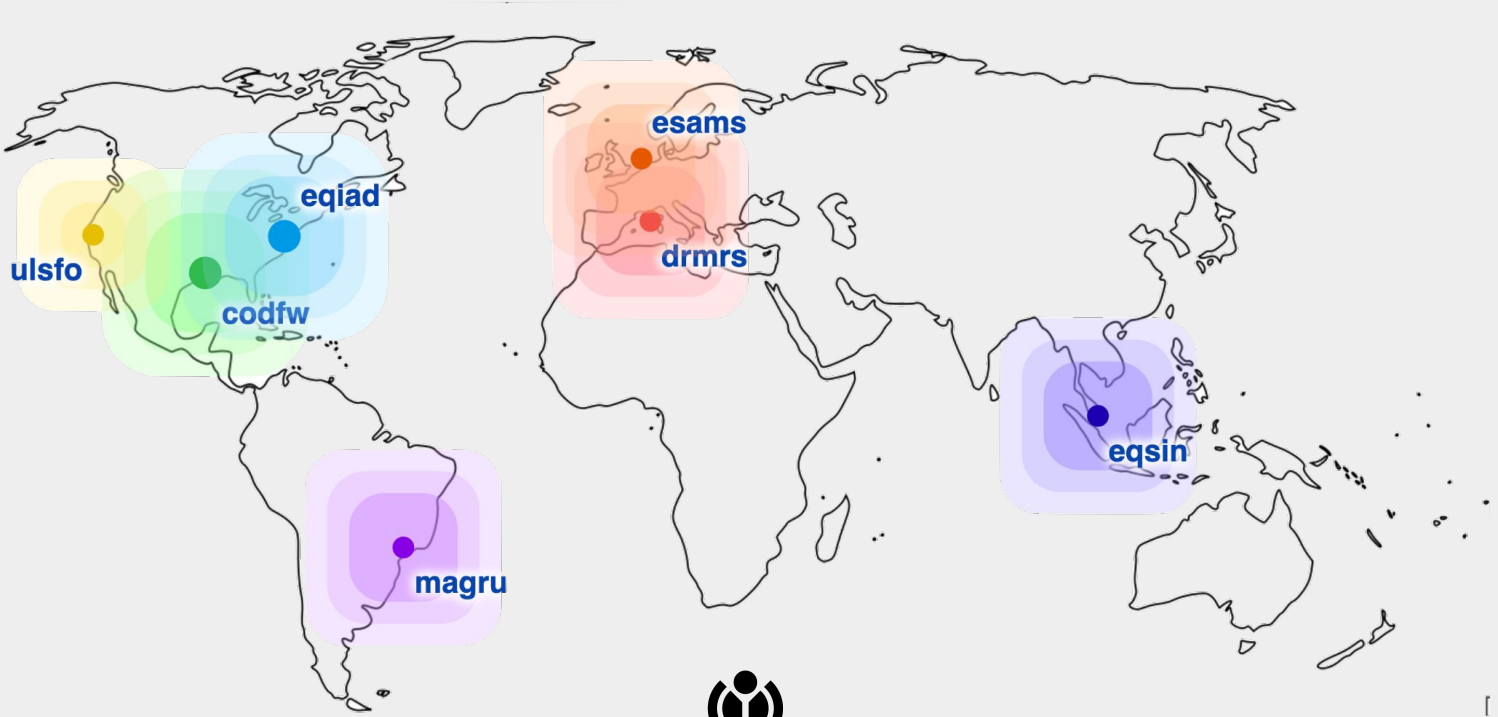


CDNs

- Speedy Delivery
  - Edge caching
  - Static content (optimised)
  - Prefetch & Pre-warm
- Routing
  - Anycast to closest node
- Reliability
  - Load Balancing
  - Less load on origin
- Security
  - TLS termination & session reuse
  - DDoS protection
    - Network Absorption
    - Application Level Protections
- Traffic Surges
  - Shields origin servers
  - Load Shedding

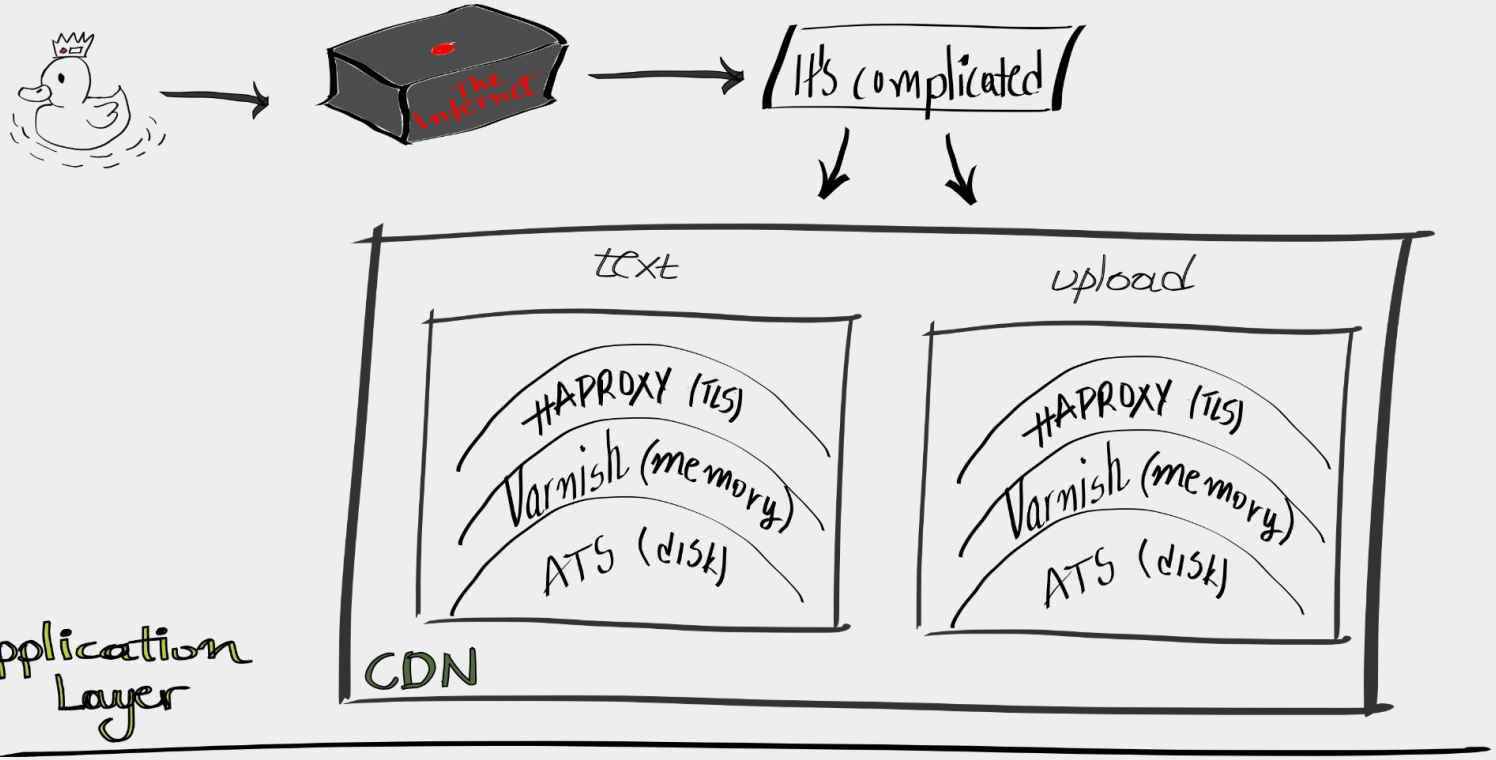
WIKIMEDIA FOUNDATION

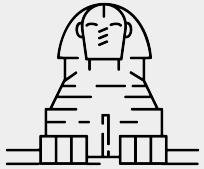
Wikimedia CDN



WIKIMEDIA FOUNDATION

What's in a cache server?





CDN Fun Facts

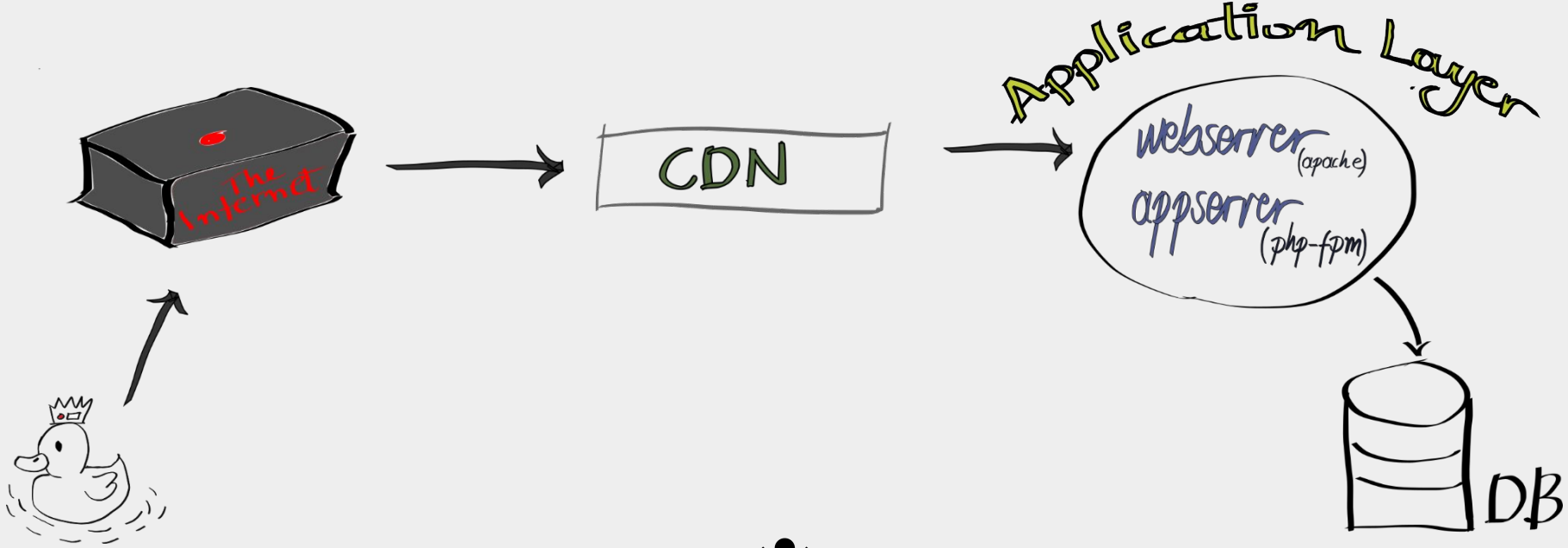
… and

Not-so-Fun Facts

- **Zipf's Law** ↠ frequency
- **Pareto distribution** ↠ 80/20
- **Regional Content Characteristics**
- **Privacy Law Compliance**
- **Licensing Requirements**
- **Censorship Regulations**

Fifty Layers of Application Caching

Basic Web Infrastructure



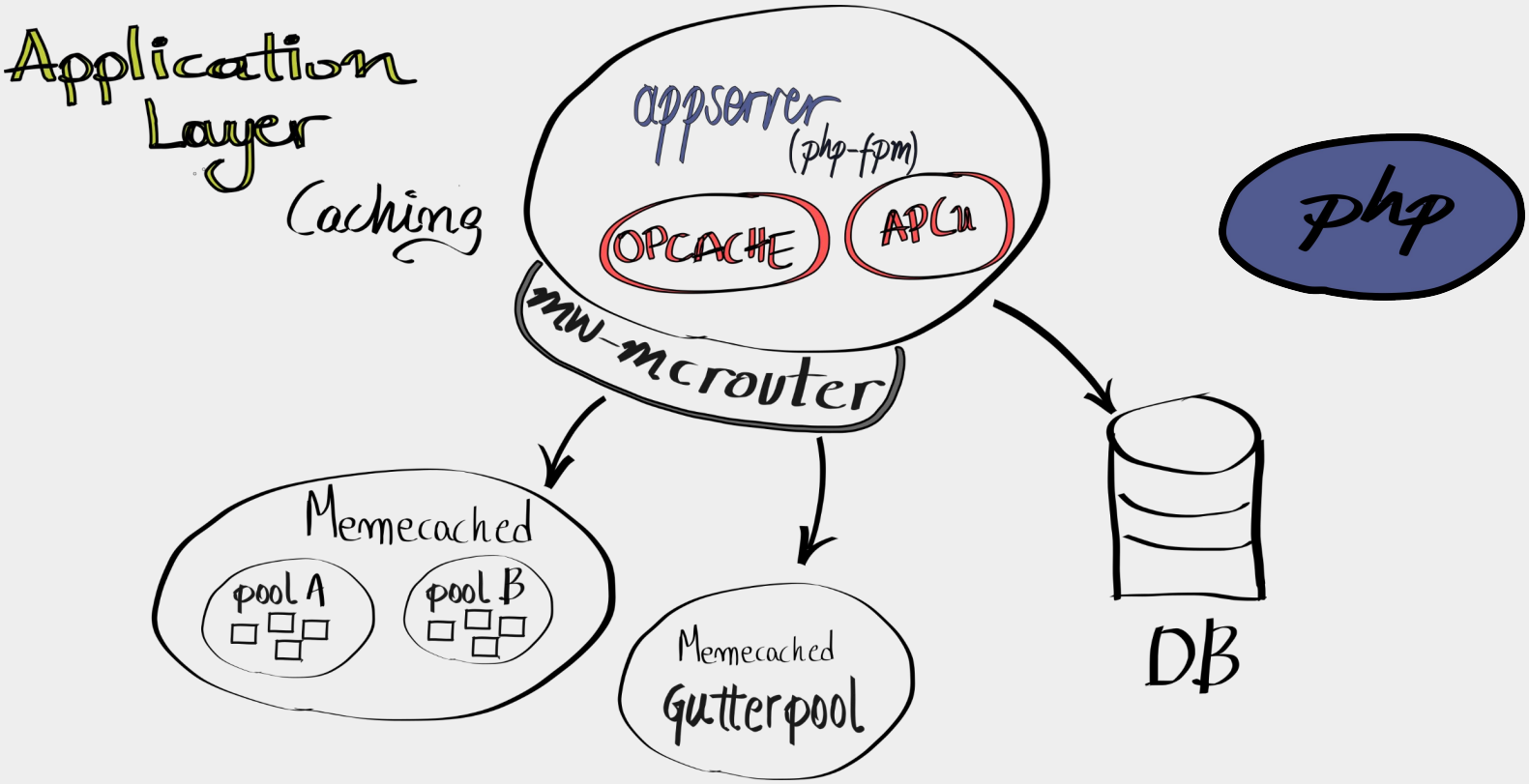
WIKIMEDIA FOUNDATION

Application Caching

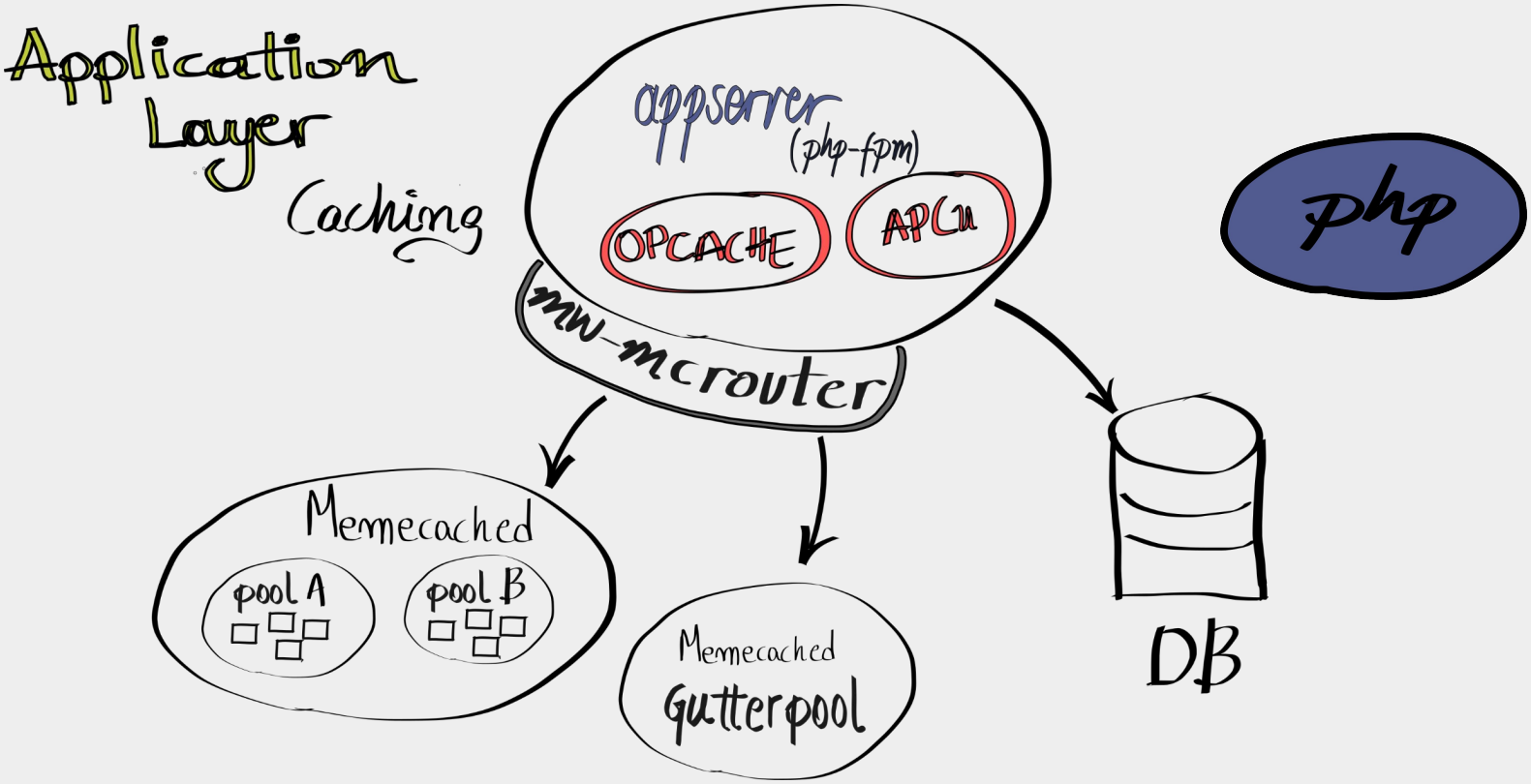
In-memory ↠ fast, limited, volatile

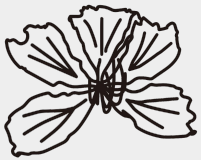
External Caches ↠ slower, scalable, semi-persistent

Application Layer Caches

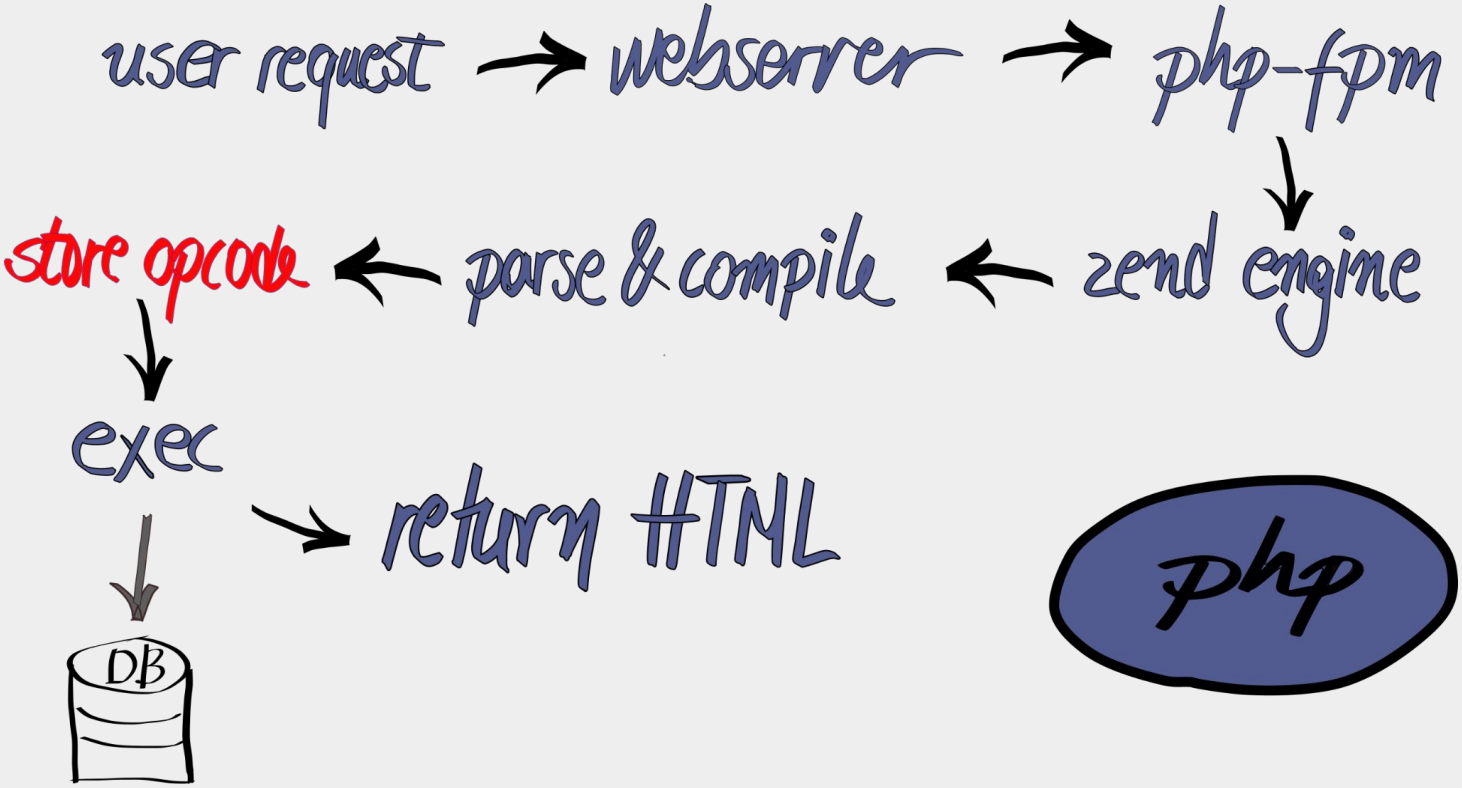


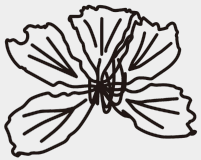
Shared Memory Extensions



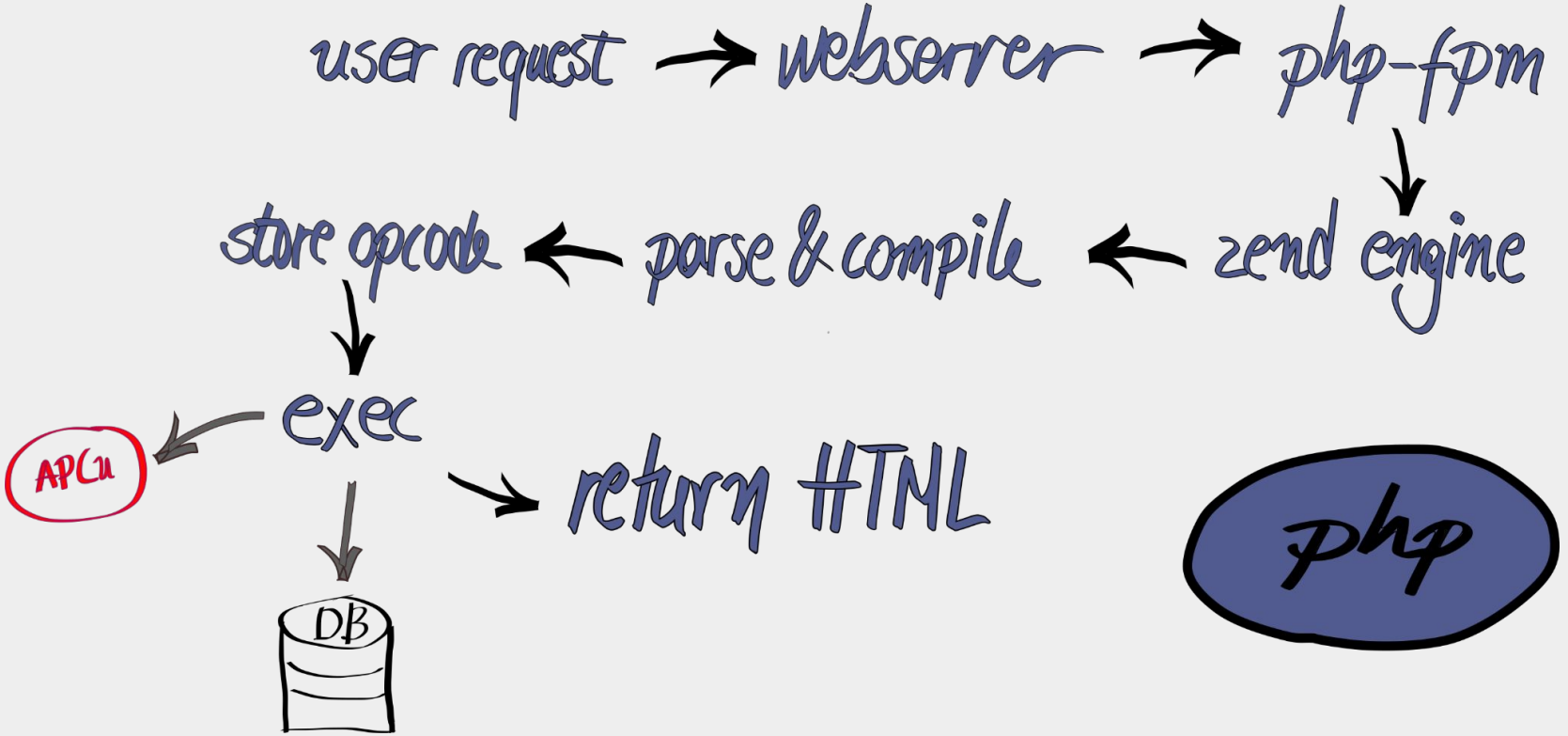


Life of PHP request





Life of PHP request



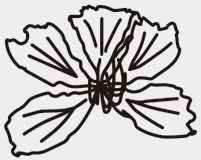
Equivalents

Java

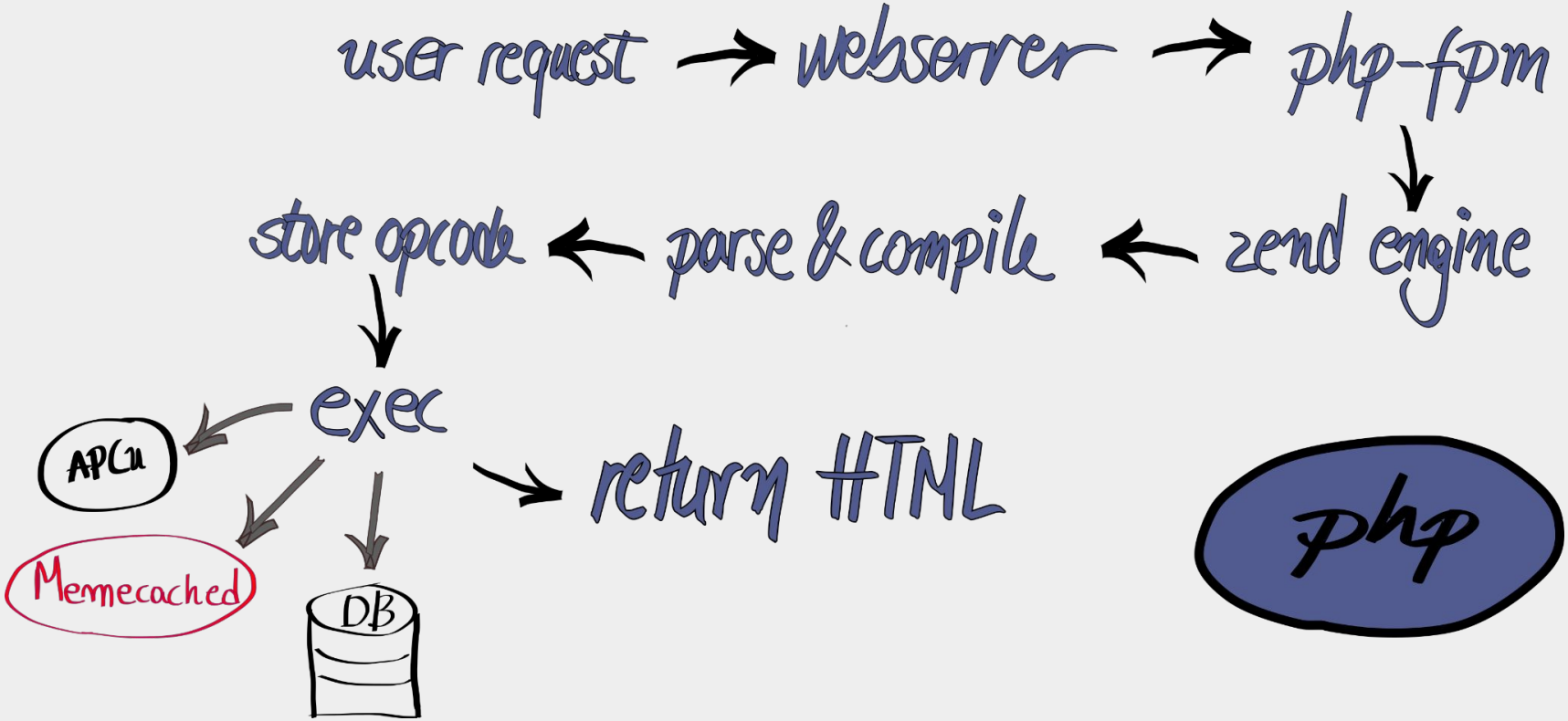
- Bytecode Caching
  - JVM JIT
  - Compiles bytecode to native machine code at runtime
- User data caching
  - Caffeine
  - Ehcache
  - Hazelcast

Ruby

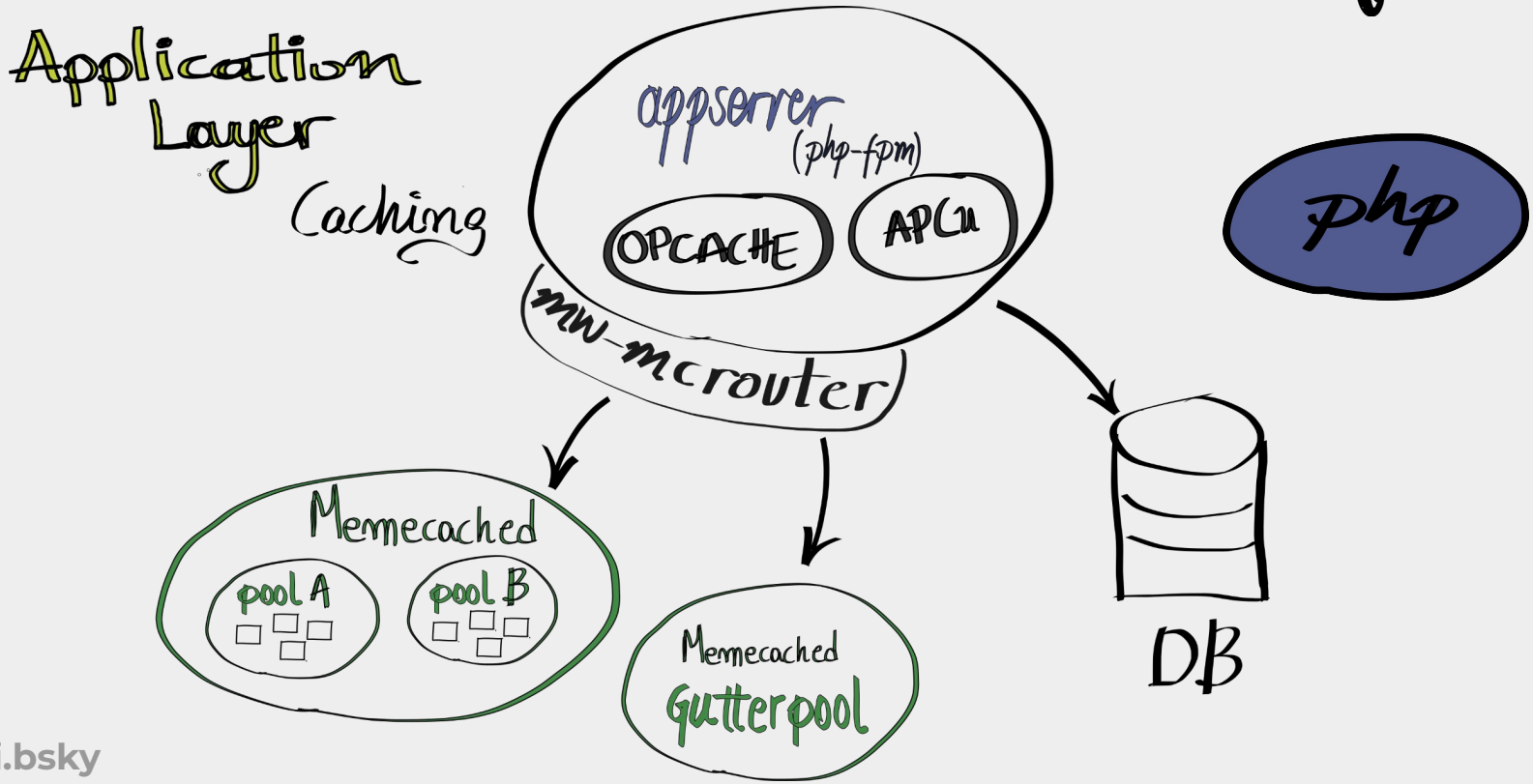
- Bytecode Caching
  - YJIT
  - MJIT
- User data caching
  - ActiveSupport::Cache::MemoryStore
  - Rails.cache



Life of PHP request



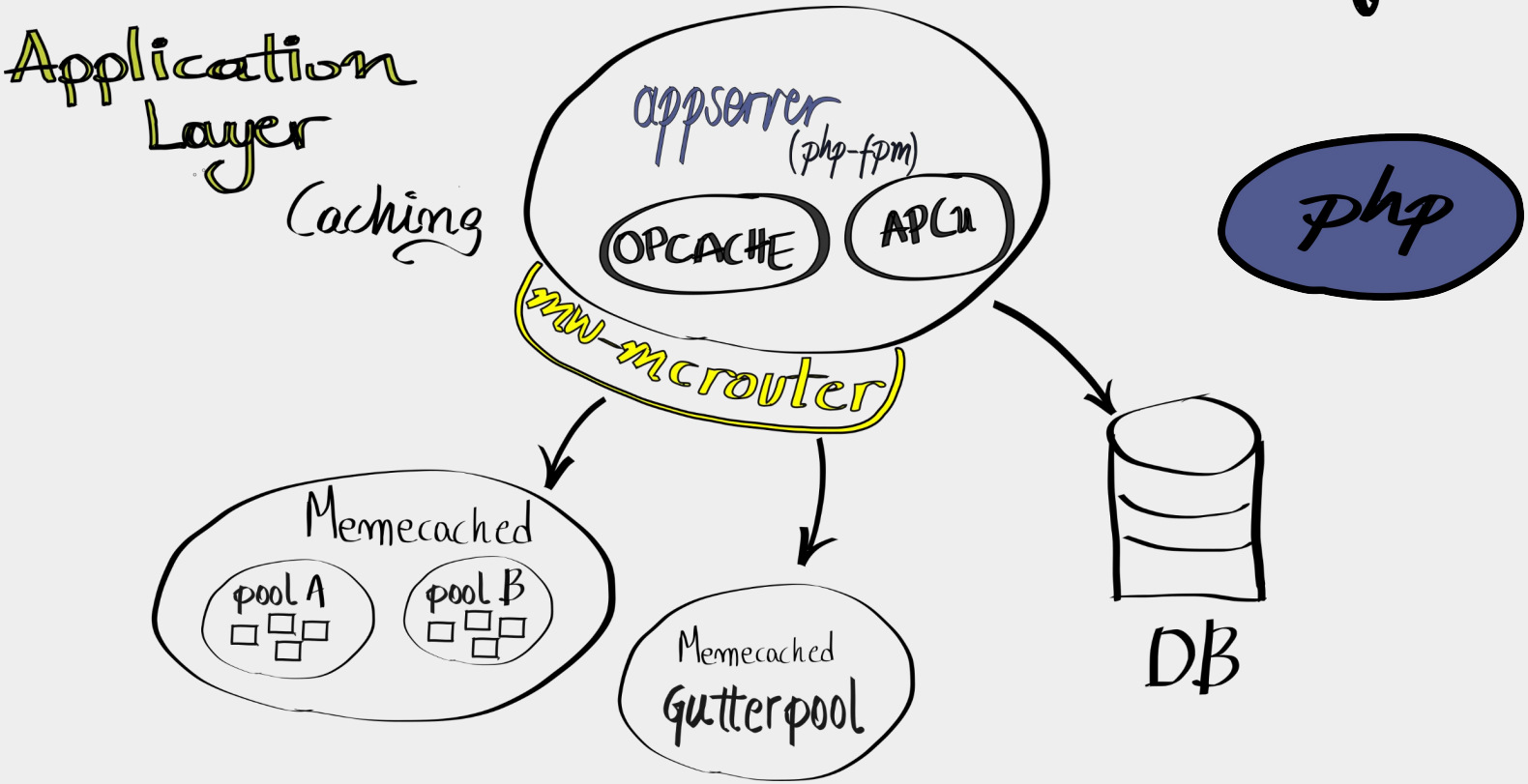
External Caches



Memcached

- Key-Value in-memory
  - Entirely on RAM
  - +disk via extstore
- Slab memory management
  - Fixed sized “Slabs”
  - Slab Classes
  - Reduces Fragmentation
- Least Recently Used
  - Hot-warm-cold
  - Tiered caching
  - Cold data pushed out first
  - Even colder in extstore

Distributed Caches



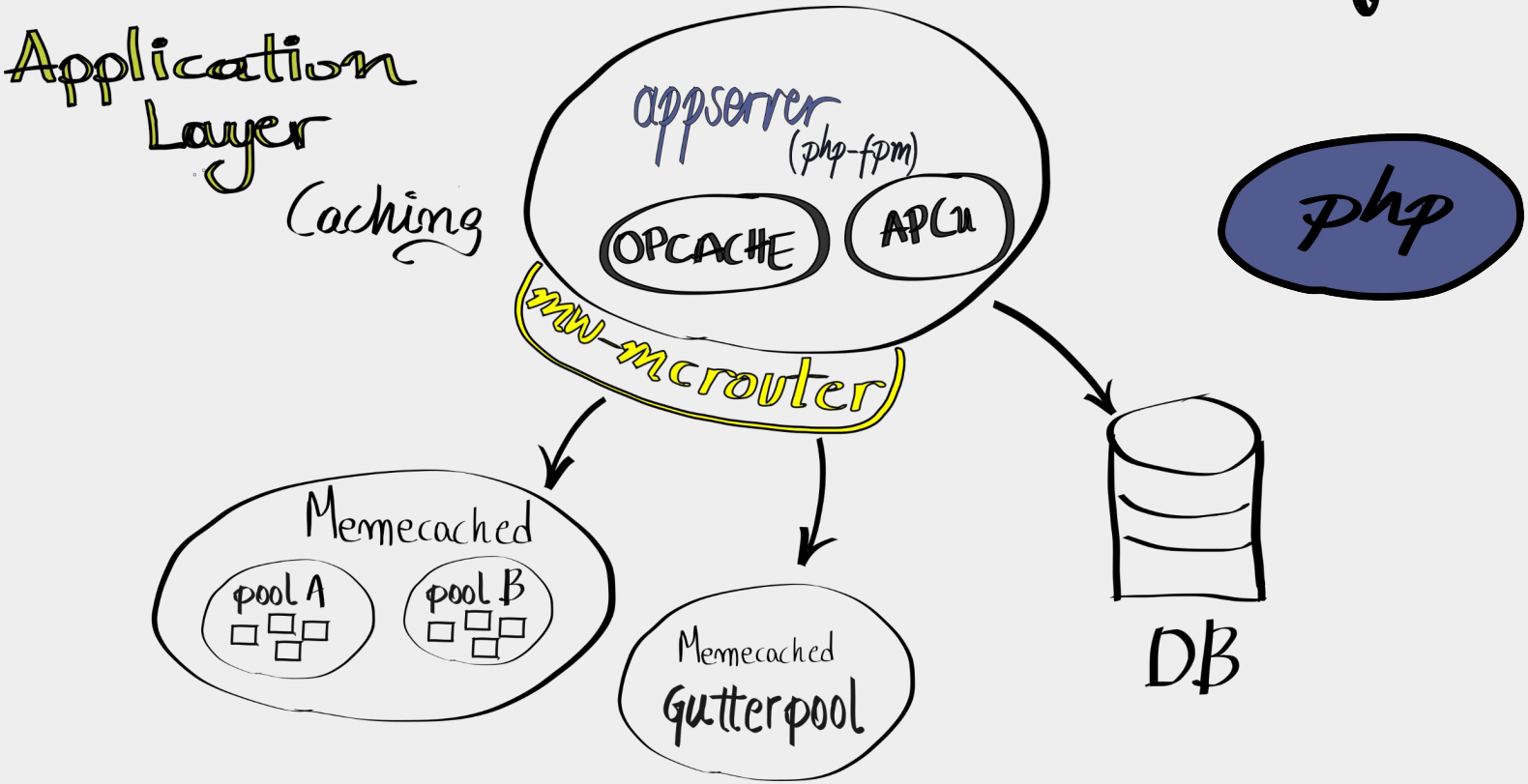
Mcrouter

- Memcached Proxy/Router
  - Open Source by Facebook
  - Transparent

Mcrouter

- Memcached Proxy/Router
  - Open Source by Facebook
  - Transparent
- Advanced Routing
  - Sharding and Replication
  - Failover and Load Balancing
  - Server Pools
  - Key Prefix routing
  - Hash based prefixed
- Pooling and Tiering
  - Efficient Connection pooling
  - Support for tiered cache hierarchies

Distributed Caching

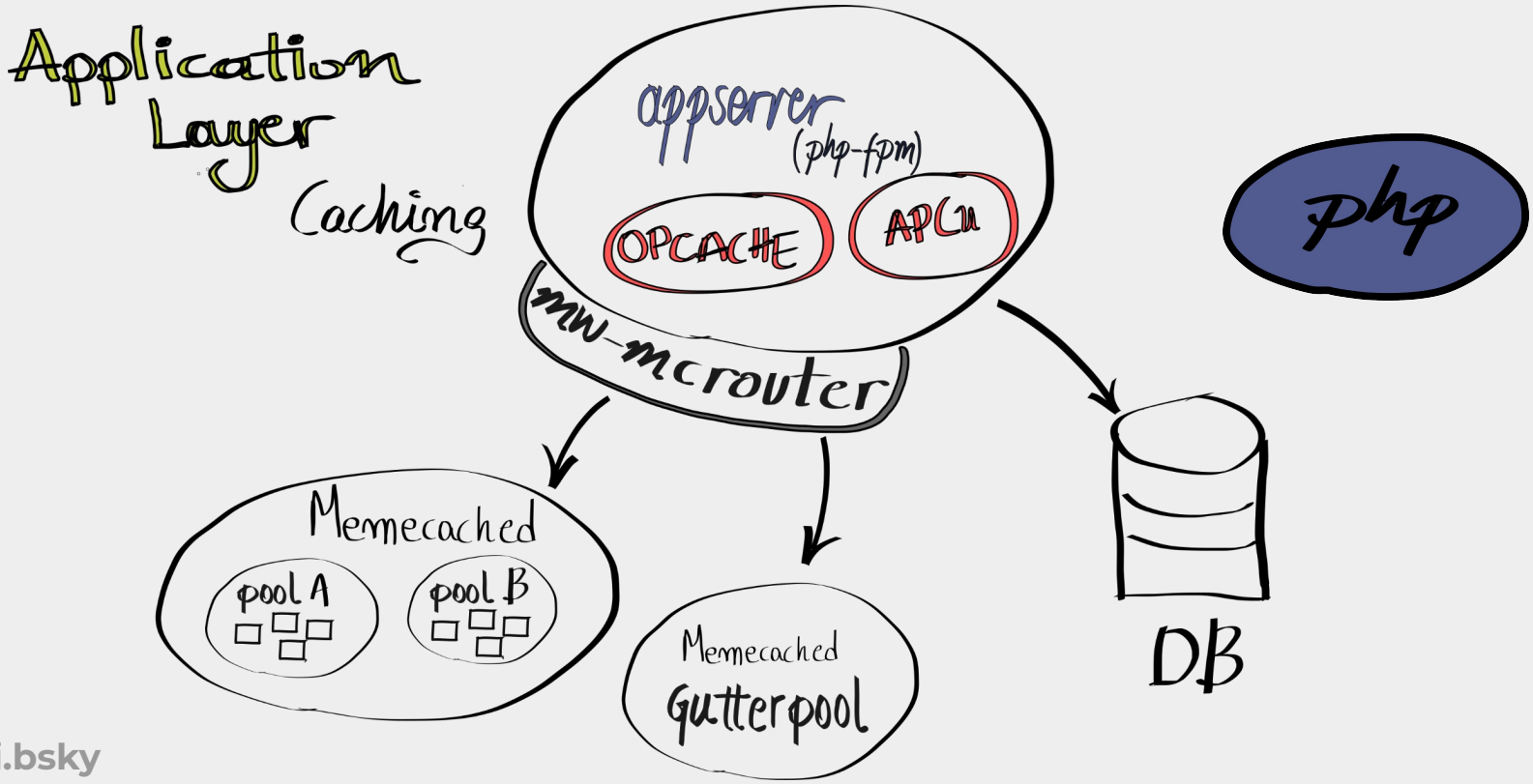


Mcrouter

- Memcached Proxy/Router
  - Open Source by Facebook
  - Transparent
- Advanced Routing
  - Sharding and Replication
  - Failover and Load Balancing
  - Server Pools
  - Key Prefix routing
  - Hash based prefixed
- Pooling and Tiering
  - Efficient Connection pooling
  - Support for tiered cache hierarchies
- Battle Tested
  - Horizontal Scaling
  - Millions of requests per second

WIKIMEDIA FOUNDATION

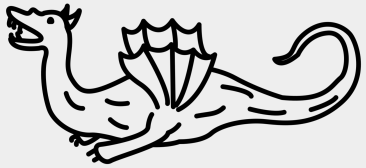
Application Layer Caches



Fifty Shades of Stale

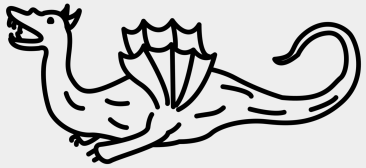
WIKIMEDIA
FOUNDATION

Warm Up



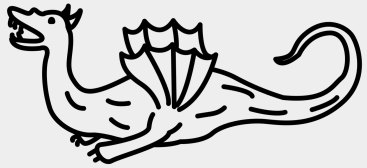
- Lazy Loading
  - On cache miss, fetches data from origin

Warm Up



- Lazy Loading
  - On cache miss, fetches data from origin
- Write-through
  - Data updated on the backed are updated to cache as synchronous

Warm Up

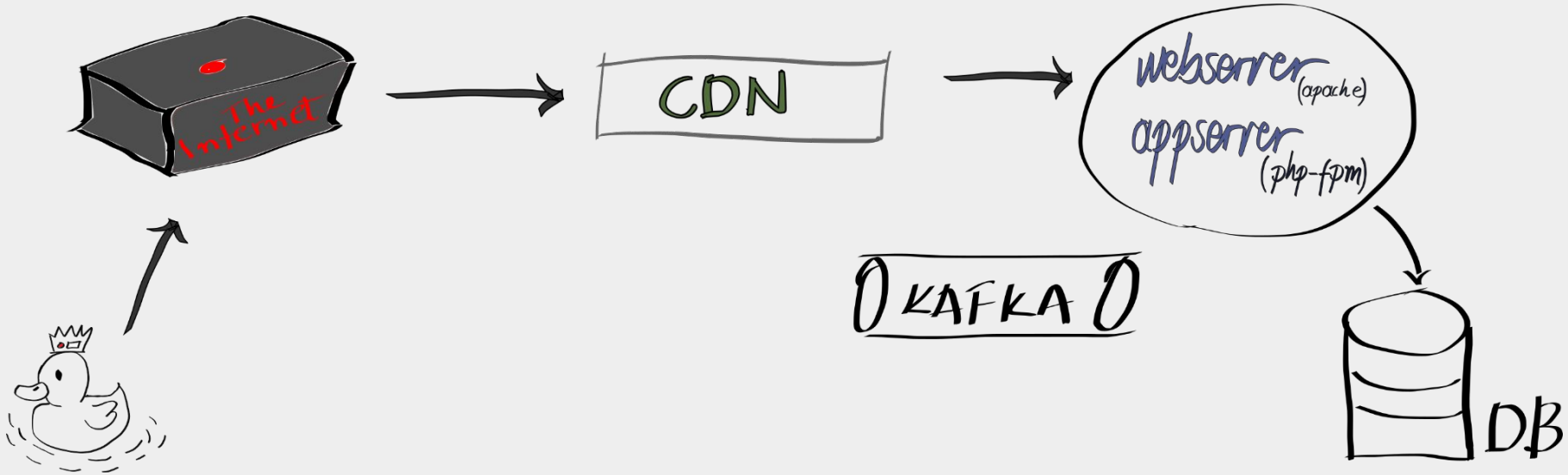


- Lazy Loading
  - On cache miss, fetches data from origin
- Write-through
  - Data updated on the backed are updated to cache as synchronous
- Pre-warm
  - Items are proactively refreshed before expiring
  - Items are proactively refreshed upon changes

Invalidate

- Time-to-Live
  - Expire after a fixed time
- Write-around
  - Data is updated only on the backend
  - Cache gets updated on next read
- Write-back
  - Write to cache first
  - Asynchronously update backend
- Manual
  - Explicitly clear caches
  - API calls
- Versioning
  - Always check the current version of the key
  - Fetch that version or recompute and store
- Event-driven
  - App emits events to invalidate caches
  - Multi-tiered caches (eg CDNs)

CDNs



WIKIMEDIA FOUNDATION

TL;DR

Browsers and CDNs cache responses for URLs.

Application-layer caches store the data needed to build those responses.

Rage Against the Machine Learning

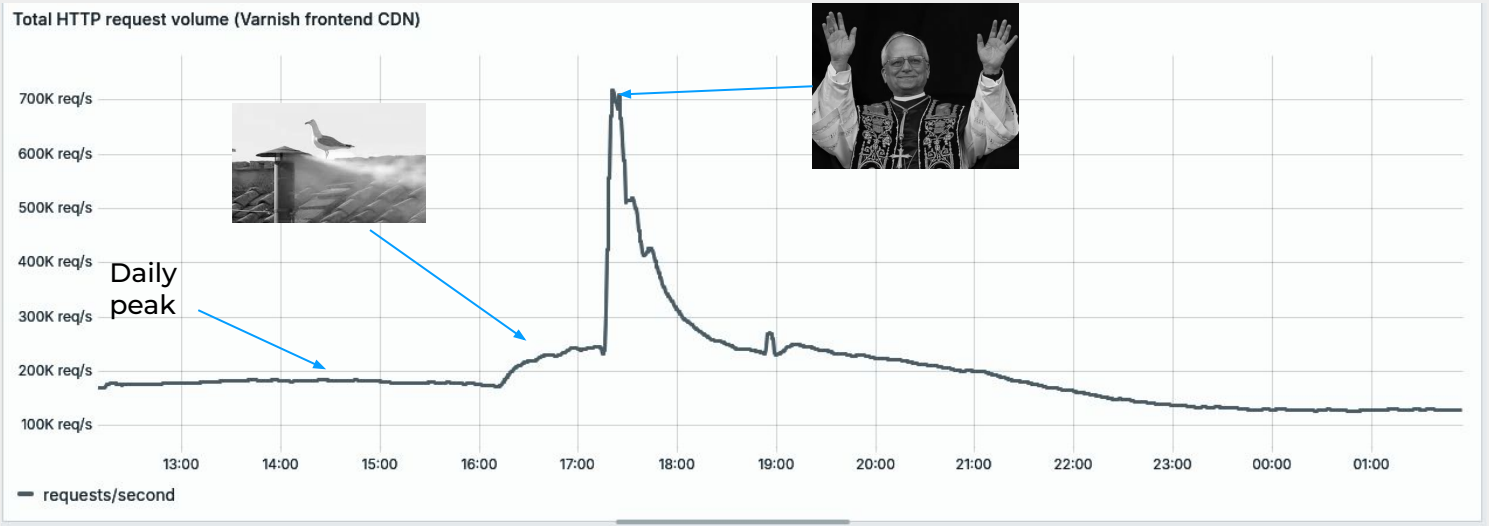
Credits: Giuseppe Lavagetto & Chris Danis

WIKIMEDIA
FOUNDATION

Optimised for humans

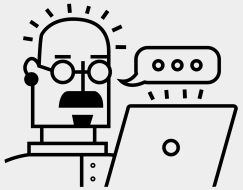
- Organic traffic naturally warms and sustains the caches
- Users scroll down, not across

Optimised for Humans



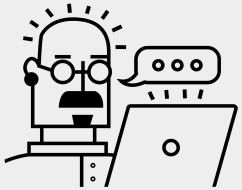
Optimised for humans

- Organic traffic naturally warms and sustains the caches
- Users scroll down, not across
- Our systems are engineered so to bend but not break

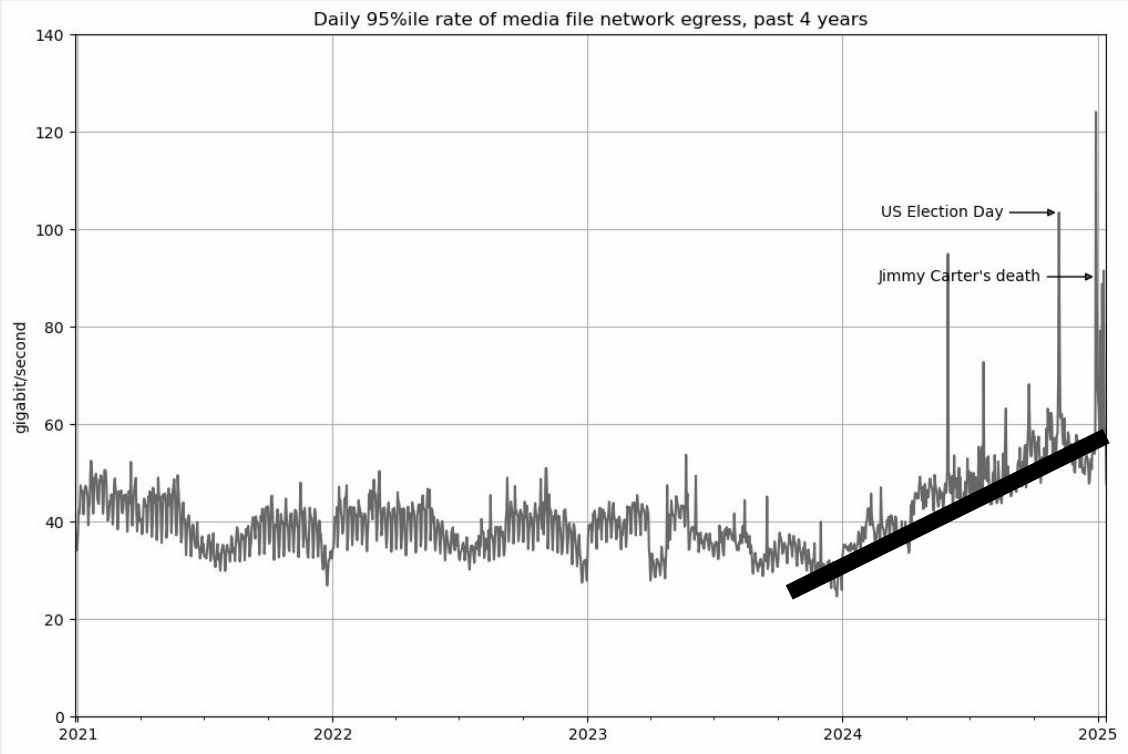


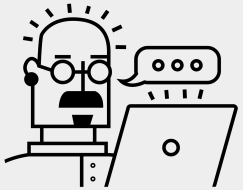
The rise of the machines

- LLM crawlers and AI agents increase global web traffic
  - Rapidly increasing web traffic
  - Not just html
  - Media with detailed metadata



The rise of the machines





The rise of the machines

- LLM crawlers and AI agents increase global web traffic
  - Rapidly increasing web traffic
  - Not just html
  - Media with detailed metadata
- Don’t play by the Rules
  - Ignore robots.txt
  - Distributed
  - Non predictable

WIKIMEDIA FOUNDATION

- They are impersonators
  - Try to impersonate other bots
    - … or other LLMs/agents
    - … or humans
    - … so residential proxies are the new black

We Pay The Bill

- Traffic costs
- Computational costs
- Power consumption
- Engineering time and monitoring fatigue

Everyone Pays The Bill

- Articles, blogs, media, news
- Code review platforms
- Even testing systems
- Our users

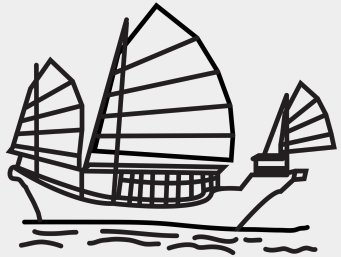
WIKIMEDIA FOUNDATION


Welcome to
CACHE
CREEK

- Dynamic Responses & personalised content
- Multiple page variants
- PBs of media
- Cache pollution
- Invalidation & warm up cost

Cache everything then?

Can we strike back?



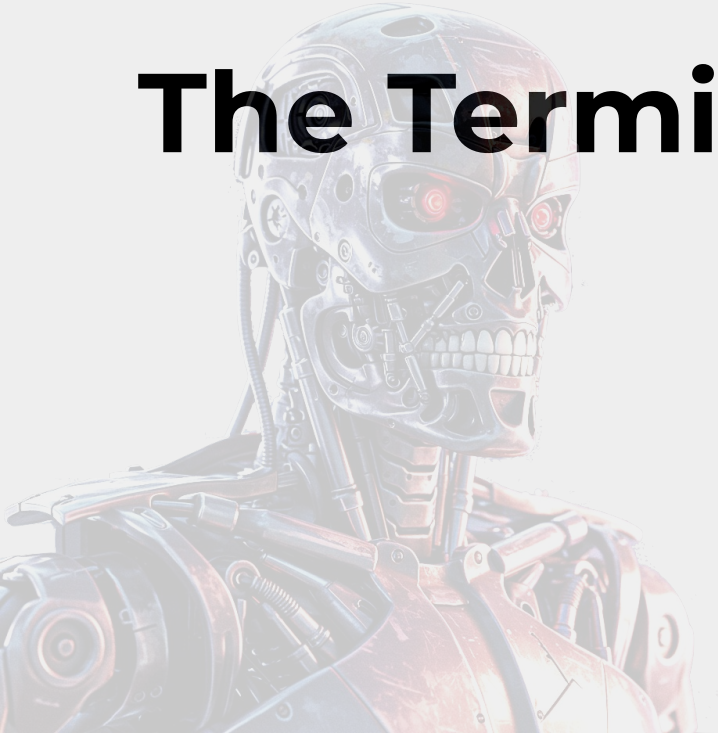
- Dumb Crawlers
  - Ban old versions of UAs, OSs etc
- Traffic Categorisation
  - IP reputation
  - Limit bw and rps
  - Block known Cloud
  - Database of crawlers
- Update robots.txt
  - One can only hope
  - Ban whatever does not respect it
- Browser checks
  - Cookies, JS, Headers
  - JA3/JA4 TLS/HTTPS Fingerprinting

WIKIMEDIA FOUNDATION

All Tied Together

WIKIMEDIA
FOUNDATION

The Terminator Was Wrong.



SKYNET
NEURAL NET-BASED ARTIFICIAL INTELLIGENCE
CYBERDYNE SYSTEMS CORPORATION
FOUNDATION

Skynet is going to scrape our websites till we run out of resources.

— the last human webmaster

Skynet vs Skynet

vs Skynet

SKYNET
NEURAL NET-BASED ARTIFICIAL INTELLIGENCE
CYBERDYNE SYSTEMS CORPORATION
FOUNDATION

Thank you!

SREcon25 EMEA
Dublin 2025
fosstodon.org/@manjiki

Links

- https://erinfeinberg.com/king-for-a-day
- https://commons.wikimedia.org/wiki/File:Welcome_sign_-_Cache_Creek,_British_Columbia,_Canada_-_July_1990.jpg
- https://commons.wikimedia.org/wiki/File:TokyoTelemessage_PHOENIX-fw_1.jpg
- https://commons.wikimedia.org/wiki/Category:Wikipedia_15_marks_in_English
- https://memcached.org/

SREcon25 EMEA
Dublin 2025

meta.wikimedia.org/wiki/User:Effie_Mouzeli_(WMF)

WIKIMEDIA FOUNDATION