A report on the decentralized web

Master's Thesis submitted to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Master of Science in Informatics
Specialization in Computer Systems

presented by
Eric Botter

under the supervision of
Prof. Fernando Pedone
co-supervised by
Leandro Pacheco De Sousa

June 2018

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Eric Botter
Lugano, 20 June 2018

Someone said . . .

Someone

Abstract

This is a very abstract abstract.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et

vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Acknowledgements

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras

ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Contents

Contents **xi**

1 Introduction **1**

2 Background: The Current Web, Models and Definitions **3**
2.1 Problems in the current web 4
2.2 A Distributed Environment 5
2.3 Decentralization trade-offs 6

3 Decentralizing Storage **7**
3.1 BitTorrent 7
3.1.1 DHT 7
3.2 IPFS 7
3.3 Ethereum Swarm 7
3.4 Filecoin 7
3.5 Analysis 7

4 Decentralizing Naming **9**
4.1 Namecoin 9
4.2 Ethereum Name Service 9
4.3 Analysis 9

5 Distributed Web Projects **11**
5.1 ZeroNet 11
5.2 Blockstack 11
5.3 Analysis 11

6 Upcoming projects **13**
6.1 Substratum 13
6.2 Hashgraph 13

7 Conclusions **15**

A The Domain Name System - A brief overview **17**

Glossary **19**

Bibliography **21**

Chapter 1

Introduction

Chapter 2

Background: The Current Web, Models and Definitions

The World Wide Web is probably the most popular and used service of the Internet, sometimes even confused with the Internet itself. It is very common nowadays to access the WWW (or most commonly known as simply "the Web") and browse websites from many platforms, from the typical desktop computer to the modern smartphone.

Let us define the scenario in which the Web lives. It is based on a client-server architecture, where Web servers provide objects (such as documents, images or files in general) to clients that request and display them, called *user agents* (e.g. Web browsers).

In the Web, documents and objects are identified by a Uniform Resource Locator (URL), whose most important component is the domain name: it is a human-readable label that identifies a device within the Internet. A domain name is composed by sequences of letters and symbols, separated by each other with dots. This separation is needed by the hierarchical structure of domain names, but we won't delve into the details of the Domain Name System here. More information can be found in Appendix A.

To access a website, the client has to know the domain name associated to that website. This is usually provided by the user or by services such as search engines. The domain name is resolved to an IP address by using the Domain Name System. Once obtained, the client opens a TCP connection towards that address on port 80, and starts exchanging messages using the HyperText Transfer Protocol (HTTP).

HTTP is a client-server, request-response protocol. Clients specify the details of the needed resource in the request and the server replies with the content or an error status if something went wrong (e.g. 404 Not Found). We won't explore HTTP as none of the projects that we will see rely on HTTP or any of its properties.

There are different ways to setup a website. A content creator can either setup a custom server and upload a website there, or it can rent a server (either physical or virtual) from an existing provider.

2.1 Problems in the current web

The main problem in the current Web is vulnerability to censorship. Since we have a direct relationship from domain names to websites (or from IP addresses to websites), it is relatively easy for powerful parties (including governments and ISPs) to block communications from users to a certain service. The main attacks that can be used to prevent communication towards a website are:

- Denial of Service (DoS): a large volume of requests is sent towards the targeted server, which quickly runs out of available resources (such as bandwidth, simultaneously open connections, memory or CPU). Requests can be sent from a single device, but in current days requests are typically sent from multiple sources, in order to both increase the volume of traffic and make it difficult to identify and stop the origin of the attack: this is known as Distributed Denial of Service (DDoS).
- IP address blocking: packets towards a given address or address range are blocked. This attack can be enacted by routers that exchange packets regarding the targeted IP address, which can interrupt forwarding of said packets thus preventing any sort of communication, making the server effectively disconnected from the Internet.
- DNS hijacking: by altering DNS resolutions, the domain of the targeted website can either be deleted or edited to make it refer to another IP address, thus preventing access to the original content. This attack can be carried out by both the owners of the DNS resolver (by directly editing their records), or by third parties through an attack called DNS cache poisoning: an attacker pretending to be a valid name server intercepts DNS requests from other name servers and provides fake responses to alter the address of the targeted domain, also setting a high time-to-live so that the redirection is active for as long as possible. Another vector for DNS hijacking, though unrelated to DNS itself, is to remotely edit the configuration of typical home routers through known vulnerabilities, changing the DNS resolver to a malicious one.

We also have a problem of trust. When you access a website, there is no guarantee that the data you received is from the content creator, because HTTP is vulnerable to man-in-the-middle attacks. There is no mechanism to verify the authenticity of the transmitted data and the protocol does not use encryption, so anyone can forge a valid HTTP communication (even based on an ongoing one) and send it through the wires. We expand on this in Section 2.2.

HTTPS resolves this issue by asymmetrically encrypting the communication channel, and authenticating the data that is sent, but the current trust system (X.509) is based on certificate authorities and is considered weak, which might allow for identity theft.

Another important issue is privacy and handling of personal information: with the current scenario, whenever you connect to a website, that website privately stores data about you. This data can be either automatically collected from user interactions, or can be provided directly by the user: consider, as an example, a social network, where users provide personal information such as their generalities, and the website collects data such as post interactions, number and timestamps of logins, and so on. This effectively moves ownership of the data from the user to the company. Data that intrinsically belongs to the user (especially personal information such as name, address and phone number) are stored privately into the company server, and the user has limited control over it, since the only possible actions on the data are the ones defined by the company or required by law.

2.2 A Distributed Environment

We have to rethink the Web if we want to move it to a decentralized environment. The current Web is a centralized system: each website is owned by a party that we'll define as *content creator*. The content creator owns the website and is responsible for distributing its content, either by using a self-owned and maintained web server or by publishing it to a dedicated service, known as *web hosting* service provider (there are too many services currently online to present a somewhat accurate list of examples here). When using *web server*, we will always refer to both these options, since in both cases there is always a web server that serves the website, whether it's owned by the content provider or by a company. Although it's not required, the content creator usually also obtains a domain name to associate with the website.

This system is centralized because the website is accessible only through the web server. If obtained, the domain name will always direct towards that server (even if it changes its IP address, since that's one of the main purposes of DNS).

Let us introduce a very important concept in distributed systems: **failure**, and its related models. We introduce it now to highlight the difference between a centralized environment and a decentralized (or distributed) one. The failure model in which we could place the Web is a **stopping failure model**: an entity can fail only by stopping, or *crashing*. This means that we *trust* each entity to behave as expected by the protocol, or to not function at all.[1]

This model allows Web clients (browsers) and content creators to make certain assumptions on the behavior of web servers and other components:

- Web clients assume that the content they request is returned without any modification, with its content exactly as intended by its creator;
- It is assumed that the channel through which the information is sent does not modify that data;
- Content creators assume that the website that they create is stored in the web server (and distributed) without any modification.

One could argue that this model does not accurately represent the reality. For example, a malicious third party can interfere in the communication channel and change the data that is transmitted. This is transparent to both the clients and the web server, since there is no mechanism in place to ensure that the data is *integral*[2] – and rightfully so, given that the communication channel is assumed to not alter data arbitrarily. This is known as a *man-in-the-middle* attack.

To allow this scenario, we need to weaken the model, by removing assumptions on what the communication channel is supposed to do. We now place the communication channel in a **Byzantine failure model**: it can fail not only by stopping, but also by exhibiting arbitrary behavior.

HTTPS, the more secure Web protocol, assumes this weaker scenario. It introduces encryption and authentication of transmitted data, which allows it to be transferred over *non trusted* communication channels. If the data is tampered with by a malicious communication channel, the client or the web server can detect it and react accordingly.

[1] This is an oversimplification of the model, but we don't need to give more detail than this.

[2] We mean a stronger integrity than the one provided by TCP. TCP ensures data integrity against transmission errors, which are in the order of few bits per kilobyte, but in this attack the entire TCP packet can be rewritten, allowing TCP packets to appear unaltered.

For example, if a content creator publishes a website on a web hosting service, that platform is able to delete or alter any file of that website, effectively sending to clients different information than the one intended by the author, and the clients would not notice this difference.

2.3 Decentralization trade-offs

Chapter 3

Decentralizing Storage

3.1 BitTorrent

3.1.1 DHT

3.2 IPFS

3.3 Ethereum Swarm

3.4 Filecoin

3.5 Analysis

Chapter 4

Decentralizing Naming

4.1 Namecoin

4.2 Ethereum Name Service

4.3 Analysis

Chapter 5

Distributed Web Projects

5.1 ZeroNet

5.2 Blockstack

5.3 Analysis

Chapter 6

Upcoming projects

6.1 Substratum

6.2 Hashgraph

Chapter 7

Conclusions

Appendix A

The Domain Name System - A brief overview

The Domain Name System is a decentralized naming system for devices connected to a network (including the Internet), currently defined with RFC 1034rfc [a] and RFC 1035rfc [b] and updated with successive RFCs throughout the years.

The DNS defines three components:

- The *domain name space* is a tree data structure, where nodes are identified by *labels*: labels compose the domain names in a hierarchical way, by concatenation of labels separated by dots. For example, for the domain "`www.example.com`", "`example.com`" is a child of "`com`" and "`www.example.com`" is a child of "`example.com`".
- *Name servers* are programs which store information about a subset of the domain space and references to other name servers which have information about the rest of the tree. Name servers have *authority* over the parts of the tree of which they have complete information.
- *Resolvers* are programs which receive queries from clients and respond with information extracted from the name servers. Resolvers only need to know directly just one name server to complete all possible queries: if that name server does not contain the requested information, the resolver uses its references to reach other name servers.

The domain name space has one root node, labeled with an empty string. Childs of this node are called *top-level domains*, among which we can find "`.com`" and "`.org`", and two lettered *country codes*, such as "`.ch`" and "`.it`". Currently, there are about one thousand different top-level domainsweb.

When resolving a hostname, resolvers query the root name server with the whole domain. The root name server usually replies with the address of the name server which has authority over the top-level domain of the hostname, but it also has facility to reply with the address of the actual server associated with the whole hostname. If the query has not been completed, the query is repeated with the correspondent top-level domain name server, and so on.

To reduce traffic towards the root name servers (and all other name servers), DNS resolvers implement a *caching* system: results from name servers are stored for reuse, together with a time-to-live value specified from the name servers themselves.

Glossary

- TCP
- ISP

Bibliography

RFC 1034 - Domain names - concepts and facilities. https://tools.ietf.org/html/rfc1034, a. [Retrieved 8-June-2016].

RFC 1035 - Domain names - implementation and specification. https://tools.ietf.org/html/rfc1035, b. [Retrieved 8-June-2016].

List of Internet top-level domains - Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains. [Retrieved 11-June-2016].