[mCaptcha] NGI Zero Entrust Grant Application

General Project Information

Project Name

mCaptcha

Website/wiki

https://mcaptcha.org

Abstract: Can you explain the whole project and its expected outcome(s). (you have 1200 characters)

mCaptcha is a privacy-focused, highly scalable Proof of Work(PoW) based CAPTCHA system with a focus on user experience and accessibility.

Existing CAPTCHA systems rely on human abilities to differentiate objects better than bots, which makes it impossible for persons with special(cognitive, auditory, and visual) needs to access the web.

Additionally, they rely on Internet Protocol(IP) address-based rate limiting and tracking the users across the internet to gauge and determine the possibility of a request being spam. Both IP and tracking users across the web make this mechanism inaccurate and impossible to use in anonymizing networks like Tor.

Current CAPTCHA systems are not idempotent and "predict" spam possibility. There is no clear documentation on how their prediction mechanisms work. The internet is an open technology and is an important means to access information. Libre idempotent CAPTCHAs are required to ensure discrimination-free access to information.

With mCaptcha, I'm trying to build just that.

Have you been involved with projects or organizations relevant to this project before? And if so, can you tell us a bit about your contributions? (Optional) This can help us determine if you are the right person to undertake this effort.

I am the creator of mCaptcha.

mCaptcha has been in active development since March 2021. Over the year, I have spent more than 1,000 hours on project writing code, documentation, project planning, deploying infrastructure and providing support to switch to and integrate mCaptcha.

Codeberg, based on Gitea, is the first major service that will adopt mCaptcha to solve accessibility issues and replace their current CAPTCHA. I will work with both Codeberg and Gitea maintainers to integrate mCaptcha in Gitea. This work has just started. This first production deployment is a major opportunity for mCaptcha and will provide proof of its accessibility claims.

Aside from mCaptcha, I actively work on ForgeFlux, a project that is involved in software forge federation and Hostea, which is a project trying to offer 100% libre, managed Gitea hosting. I'm a strong believer of Free(as in freedom) Software, so all my work is exclusively licensed under Free Software licenses.

Internet privacy is not accessible to everyone, but I'm in a privileged position to self-host services, which are available for anyone to use free of charge at batsense.net.

Requested Support

Requested Amount (between 5,000 and 50,000 euro)

26,200

Explain what the requested budget will be used for? Does the project have other funding sources, both past and present? (If you want, you can in addition attach a budget at the bottom of the form) Explain costs for hardware, human labor (including rates used), travel cost to technical meetings, etc.

The project has never had any funding sources.

- Developer pay: I am currently the only developer involved in mCaptcha. I require 2,000 EUR per month to sustain my current lifestyle in India. So that's 24,000 EUR per year.
- 3,500 EUR for a build server: mCaptcha requires extensive testing and must support a variety of platforms. Dedicated build server is needed to reduce build times (right now 30 mins. per build on Github CI).
- 1000 EUR for cloud rent:
  - Run experiments to determine mCaptcha's effectiveness against DDoS attacks. The source code for this experiment exists at mCaptcha/dos but needs to be run in with multiple attacking nodes to accurately simulate DDoS attacks
  - Run a survey to aggregate PoW performance on a wide range of devices: The program for this survey exists and is available at survey.mcaptcha.org and is running out of my bedroom. I would like to move it to a more stable environment before I start the campaign and circulate the survey link.

Compare your own project with existing or historical efforts. (e.g. what is new, more thorough, or otherwise different)

- Friendly captcha: PoW-based CAPTCHA but non-free. A libre, self-hostable version is available but is stripped down and doesn't have critical features like variable difficulty scaling(which improves UX). Little to no documentation on how the CAPTCHA system works.
- pow-captcha: PoW based CAPTCHA libre implementation. Doesn't support variable difficulty scaling. Uses Scrypt against SHA256, not suitable for large-scale deployments as Scrypt is more resource intensive. Integration libraries and client libraries don't exist for popular stacks yet.

What are significant technical challenges you expect to solve during the project, if any? (optional but recommended)

Objective 1: Proof of Work accessibility: PoW within mCaptcha is configurable to take anywhere between 200ms to +10s to compute. But the time will vary across devices, as some might have more computing power than others. Older devices might be too slow to generate PoW and fail validation.

I'm working on a survey to benchmark PoW on various devices, which will shed light on optimal PoW difficulty configuration. The survey results will help mCaptcha admins pick a difficulty factor that works for most visitors and improve PoW accessibility.

Activity 1.1: Run WASM and JavaScript polyfill benchmarks Activity 1.2: Collect statistics Activity 1.3: Periodically publish results and determine 90th percentile and 99th percentile average

Objective 2: Horizontal scaling to handle websites with large visitors: mCaptcha implements actor model in which each website is an actor. While the current system can scale horizontally to accommodate more websites, it doesn't scale horizontally to accommodate more visitors on a single website. If popular services are to switch to mCaptcha, single website horizontal scalability should be implemented.

Activity 2.1: Create distributed cache that implements a leaky bucket algorithm Activity 2.2: Verify correctness of the implementation Activity 2.3: Create multi-node DDoS attack simulation for benchmark(mCaptcha/dos is WIP, needs Infrastructure as Code to deploy multiple attacking nodes) Activity 2.4: Benchmark and compare performance against the semi-distributed, Redis-based cache implementation that is currently used

Objective 3: Integration tests: While there is extensive unit testing(+90% code coverage on all repositories), there are no integration tests that span all repositories.

Activity 3.1: Create Selenium-based integration testing with the following configuration options: 1. PoW options: i. WASM library ii. JavaScript polyfill library 2. Database options: i. Postgres ii. MariaDB 3. Cache options: i. embedded ii. Redis-based(uses mCaptcha/cache, a custom Redis-module) iii. Custom distributed cache 4. On browsers: i. Firefox ii. Chromium

Describe the ecosystem of the project, and how you will engage with relevant actors and promote the outcomes?

I hope to make Codeberg my showcase, and will approach other FOSS projects with non-optimal CAPTCHAs or proprietary ones with it, as well as with performance results and findings from the research.

And since mCaptcha is being deployed in Codeberg to improve accessibility, people with special needs on the Fediverse will be both the project's testers and advocates.

Attachments

None